Mind Ascribed

An elaboration and defence of interpretivism

Bruno Mölder



Advances in Consciousness Research

80

John Benjamins Publishing Company

Mind Ascribed

Advances in Consciousness Research (AiCR)

Provides a forum for scholars from different scientific disciplines and fields of knowledge who study consciousness in its multifaceted aspects. Thus the Series includes (but is not limited to) the various areas of cognitive science, including cognitive psychology, brain science, philosophy and linguistics. The orientation of the series is toward developing new interdisciplinary and integrative approaches for the investigation, description and theory of consciousness, as well as the practical consequences of this research for the individual in society.
From 1999 the Series consists of two subseries that cover the most important types of contributions to consciousness studies:
Series A: Theory and Method. Contributions to the development of theory and method in the study of consciousness; Series B: Research in Progress. Experimental, descriptive and clinical research in consciousness.
This book is a contribution to Series A.

Volume 80

Mind Ascribed. An elaboration and defence of interpretivism
by Bruno Mölder

Mind Ascribed

An elaboration and defence of interpretivism

Bruno Mölder
University of Tartu

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Mölder, Bruno.
Mind ascribed : an elaboration and defence of interpretivism / Bruno Mölder.
p. cm. (Advances in Consciousness Research, ISSN 1381-589X ; v. 80)
Includes bibliographical references and index.
1. Philosophy of mind. I. Title.
BD418.3.M635 2010
128'.2--dc22 2010028101
ISBN 978 90 272 5216 6 (Hb ; alk. paper)
ISBN 978 90 272 8774 8 (Eb)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

For my parents

Table of contents

Part II. Elaborating and defending the ascription theory

Part III. Extending the view

Acknowledgements

This book has been in gestation for almost a decade. I was hooked by the interpretivist approach already in 1997, when reading Dennett's papers. There seemed something deeply right about it, but at the same time there remained a nagging feeling that it is not altogether straightforward to pin down what exactly is right about interpretivism and how should this position be formulated and understood. I hope that with the present book I have managed to soothe this feeling.

This book stems from my University of Konstanz dissertation, refereed by Prof. Dr. Wolfgang Spohn and Prof. Dr. Jürgen Mittelstraß. The oral examination took place in Konstanz at the 10th of May 2007. I am grateful to my referees for their kind remarks. When working on these issues I was supervised by Wolfgang Spohn and Wolfgang Benkewitz and I thank both for their guidance and encouragement. My studies in Konstanz were supported by a scholarship from Herbert Quandt Stiftung and a three-year scholarship from DAAD.

My thinking on interpretivism has benefited from discussions and advice from several people in different periods. The earlier developments of the ascription theory appeared already in my University of Tartu MA thesis in 1999 and in my MPhil thesis at the University of Cambridge at 2000. I am grateful to all who helped me in those early days, especially to Ross Harrison for his supervision sessions.

The bulk of the present work was finished in the summer of 2006, after that its content has not changed significantly, although I have made several revisions in order to increase the readability and to make the presentation more concise and concentrated. In this regard, I would like to thank Daniel Cohnitz for helpful comments, Margit Sutrop for organisational support and Paul McLaughlin for checking my English. Special thanks to the series editor Maxim Stamenov and to an anonymous reviewer for comments and suggestions. Finally, I am grateful to my parents, Käthe and Olev Mölder, for their moral support over the years. This book is dedicated to them.

While working with the manuscript in its latter stages, I acknowledge the support by the basic financing scheme "Information and representation", the targeted financing scheme No. SF0180110s08 at the University of Tartu and the Estonian Science Foundation grant No. ETF7163.

The trains of thought developed in this book were presented in various occasions and I am thankful to the respective audiences: the seminar on truth in Järveküla (May 1997); University of Tartu (June 1999 and December 2002); Logic in Philosophy workgroup colloquium at the University of Konstanz (January 2002); NAMICONA workshop in Hotel Koldingfjord (December 2004); The first annual Estonian philosophy conference at the University of Tartu (May 2005); Turku-Tartu philosophy workshop (June 2005); University of Latvia (May 2006); GAP.6. conference in Freie Universität Berlin (September 2006); Toward a Science of Consciousness 2007 conference in Budapest (July 2007).

A earlier version of Section 6.2 was published as "An argument for interpretivism" in *Selected Papers Contributed to the Sections of GAP.6: Sixth International Congress of the Society for Analytical Philosophy*, Helen Bohse & Sven Walter (eds), Paderborn: Mentis, 2008, pp. 474–483.

CHAPTER 1

Preliminaries

If one happens to leaf through various current introductory treatments of the philosophy of mind, then one cannot fail to notice that beside dualism, short shrift is usually given also to various interpretivist positions, if they are treated at all. Functionalism and various reductive positions receive far more extended treatment. This, I find, is not insignificant. It says something about the present-day philosophy of mind. It says that in general, interpretivism – in very broad outline, the view that having mental properties is a matter of interpretation – is considered as not a very viable option, perhaps worth knowing about and being warned against, but not compelling enough to be defended and elaborated at length. However, I think that this quick dismissal is unfortunate for the philosophy of mind as a whole, for it leads to a distorted picture of the mind as well as the tasks of the philosopher who studies it. The main purpose of this book is to develop a version of interpretivism, the ascription theory, which, as I hope to show, is able to accomplish much of what is expected from a philosophical account of mind. I shall thus attempt to present and defend an overarching account of mind and to show how it can answer a wide range of philosophical challenges. When I apply an interpretivist approach to such topics as perception, self-knowledge and mental causation in the later chapters, my main aim is to provide a fresh perspective on these topics or just to show that my particular version of the view can accommodate several issues that have been generally considered insurmountable for interpretivism.

In this introduction, I give an outline of the book and explicate some background assumptions that are presupposed in the following chapters.

1.1 Mind and content: First steps

What is at issue in this book? In general, it is nothing less than the status of the mind in the world. In short, the question is: how should we understand the mind and mental content? I will propound a general account of what it is to have a mind in the form of explaining what it is to have states with mental content. In this way we will gain an unified account that gives a similar answer to the

questions concerning the possession of mental states and the possession of mental contents. The answer is, in general, that having a mind is being subject to a certain kind of interpretation.

What is mind? Apart from some substance-dualists, most philosophers would agree with the following specification: a subject's mind is nothing separate from the bundle or totality of mental states, events, and properties of the subject. The widespread contention is to approach the mind by accounting for such mental items. It is thus assumed that once one has given an account of their nature, no additional account remains to be given of the mind.

It is also quite common to assume that at least the initial characterisation of mental states is given in folk-psychological terms. That is, when talking about mental states, philosophers talk about folk notions such as beliefs and desires and thoughts. I shall follow this usage and conceive mental states as they figure in folk psychology. The version of interpretivism that will be presented here concerns itself only with mental states and properties as they are specified in folk psychology. My full treatment of folk psychology will be given in Chapter 4 and the relations between neural and common-sense explanations are discussed in Sections 2.6 and 7.5 of the book.

Folk psychology can be viewed as a collection of platitudes or a body of knowledge that guides the application of mental terms. In this regard, what is required for the application of mental terms lies out in the open, in our common-sense psychology and it can be mastered by anyone who masters folk psychology. One could resist this view by becoming an externalist about mental terms and arguing that when we use mental terms, we are, in fact, talking about something else: we are actually talking about brain states.[1] But if this reference is accessible to us only after serious empirical investigation, then it cannot guide the everyday use of mental vocabulary and then it would not be helpful in explicating how we understand and use mental concepts. As will become evident in the following, I do by no means deny that there is an interesting account to be given by various sciences. All I say is that, as far as the folk-psychological description is concerned, it makes sense to defend an interpretivist approach about the conditions in which this description can be applied.

I do not think that it is right to apply an externalist theory of reference to mental terms. Such theory rests on the presumption that mental terms are natural kind terms, the real essence of which can be discovered by science. This has the consequence that, contingent on possible scientific developments, it could turn out that our common-sense platitudes are false, but we have still succeeded in

1. Note that externalism about mental terms differs from externalism about mental content, the view that content is determined by factors external to a thinker.

refering to some kinds, the nature of which is unknown to us. It is not easy to see what the merits of such a way of speaking are. The ascription theory, by constrast, proceeds from the assumption that most mental terms do not denote natural kinds. There is not enough uniformity in the mental realm and nor are there firm distinctions between mental types to allow the existence of natural mental kinds. A similar view was espoused by Colin McGinn to strengthen the Davidsonian case against psychophysical laws. He noted that "the essences of mental states are right on the surface, comprised in our ordinary means of identification. … our understanding of psychological terms confirms, because it is in part explained by, the thesis that such terms do not denote natural kinds" (McGinn 1978: 217). I will have more to say about externalism concerning the reference of mental terms in Chapter 4, where I also reject the very idea that relying on the theory of reference would make a difference in the debate about the status of folk terms.

Similar claims can be made about the mind. Although the mind is approached through the account of mental states, there remains the question of whether the mind encompasses only mental states as they are constituted by folk psychology or whether it makes sense to say that brain scientists can make discoveries about the mind. In a sense, this is a verbal question, a question about the conditions of applying the term 'mind'. I take it that usually the general term 'mind' is applied to the totality of folk-psychologically constituted mental states. If one wants to talk about brain states, then using the word 'mind' is of little assistance. It is only appropriate in making sense of the correlation between various subpersonal states and mental descriptions.[2] However, this is an extended use of the term. I am not making prescriptions on use, but I would prefer to understand mind primarily through mental vocabulary. Nevertheless, in view of such extended usage of 'mind', the title of the present book, "Mind Ascribed", should be taken as tongue in cheek, for by this I mean only that the possession of mental states as constituted by folk psychology is dependent on ascription. The fact that one has a brain inside one's skull is obviously not a matter of ascription.

The notion of content also needs some initial elaboration. A common, dictionary-type specification says that content is "that which is expressed by an utterance or sentence: the proposition or claim made about the world" (Blackburn 1994: 79). However, even the shortest glance at the relevant literature shows that philosophers are far from being unanimous in their specifications and uses of the

---

2. The term 'subpersonal' originates in Dennett (1969: 90–96), where he distinguishes between the subpersonal level that concerns entities and explanatory constructs that make up a subject's body and the personal level, at which explanations in terms of an agent are provided. From time to time, I use the term 'subpersonal states' to pick out both neural and functional states. See also Hornsby (1997: Chapters 9, 10).

notion of content. The meaning of this notion ranges from the abstract proposition to the way of experiencing the world. An example of the first way of understanding content is provided by Christopher Peacocke, who unpacks his use of the notion thus:

> I use 'content' and 'Thought' as stylistic variants, and take Thoughts to have the four properties Frege attributed to them. Thoughts have absolute truth values, without relativization to anything else; they are composite structured entities; they are the objects of belief, intention, hope and the other attitudes, and it can be that one and the same Thought is judged, argued or agreed upon by two different thinkers. (Peacocke 1986: 1–2)

This is to take 'content' as the notational variant of 'proposition' – it is the object of a representational state or one part of the relation in '*x* represents that *p*'. In this sense, content is specified in connection with representing aspects of the world: it is *what is represented*, in contrast to the representation token itself, which can be taken as the vehicle of content. It should be stressed that if one means by 'content' that which is represented, then the specification of what is represented need not be limited to the list of worldly properties and affairs. It can also highlight the way it is represented. In Fregean terms, for example, the proposition can be taken as consisting of the senses of its components. In that case, the notion of content turns out to be more fine-grained than the purely externalist or Russellian specification of the object of representation. However, on the Fregean view, although the proposition consists of senses, it would still be objective as these senses can be grasped by different subjects.

By contrast, the notion of content that one comes across in the philosophy of psychology is more internalist in its nature – the focus has moved inside the subject's head, instead of the objective realm of *abstracta*. An example of this internalist notion of content is Kent Bach's characterisation of methodological solipsism:

> the psychologically appropriate way to individuate types of beliefs, desires, and other intentional states is by their contents, not by relations the subject has to things in the world, which could have been different without affecting content. When we cite such states to explain or predict actions, we specify the representational contents of those states. *What matters is how the subject represents the world, not how the world actually is.* (Bach 1982: 123, my italics)

The list of quotes can be continued, but it should be evident that the notion of content is not always used in the same sense and for the same purpose. Next, I shall provide a short overview of the main purposes for which this notion is used and then I will specify how I am going to approach to it. There are three main problems that philosophers have attempted to make more tractable by means of

the notion of content. These are the issues of objectivity, intentionality, and explanation. Let us discuss them in turn.

The issue of objectivity. One way to explain the success of interpersonal communication is to say that both parties grasp the same content or proposition of what is said. The Fregean (although Frege was not the first to employ the idea) notion of *Gedanke* (translated into English as 'Thought' or 'content') is simply a notion of this sort – it is abstract and it can be grasped by different subjects at different times. There are different notions of objectivity, but according to one sense, something is objective if it is in principle accessible to more than one individual. In this sense, the Fregean content is objective, for it is not reserved for the individual subject only. Of course, if we bring in the notion of content to explain interpersonal communication, this explanation can only be preliminary, giving rise to additional questions that need a substantial theory of content.

The issue of intentionality. Sometimes 'content' is taken just as another term for 'intentionality', or the aboutness of mental states. My thought 'cat' is about a cat. The problem of intentionality is the task of explaining how such aboutness is possible: what makes it the case that thought is about something? In recent decades, a naturalistic constraint is placed on this project. It is said that the explanation of intentionality has to be naturalistic. By this, it is meant that the *explanans* of intentionality may not contain supernatural or mysterious posits. On the contrary, the properties or conditions cited in the *explanans* should be such that there is no insurmountable difficulty in appreciating their place in the natural world.

The issue of explanation. If there is a naturalistic way of grasping intentionality, it can be put to use in explaining other phenomena. There are different strands that should be distinguished in respect of explanation. The issue that lies at most on the surface is the everyday fact that we use intentional vocabulary to explain and predict the behaviour of other people. This is done by making sense of their actions in terms of their mental states with contents. The explanation by content has led several philosophers to the idea that in order for this explanation to succeed, there have to be some conditions internal to the person whose action is explained. This idea resulted in the representational theory of mind. The internal conditions were initially specified in terms of mental states and processes. These were explained in turn by relations between internal mental representations and processes that involve such representations. In one version of this explanatory project, a whole internal language is postulated (see Fodor 1975). Here it is assumed that mental representations make up a "language of thought", which is explanatorily basic both to linguistic meaning and to propositional attitudes of common-sense psychology. According to this approach, having a propositional attitude is being related to a token of mental representation in one's internal language of thought.

Another explanatory issue relying on the notion of content is the attempt to explain meaning in terms of mental content. This approach, influenced by Grice (1957), set itself the task of reducing the meaning of words and sentences to mental content. This was supposed to be accomplished in two steps. First, one attempts to explicate the speaker-meaning of utterances in terms of the communicative intentions of speakers. Second, one attempts to reduce any residual part of the meaning (the expression-meaning) to the speaker-meaning. At present, the Gricean approach is relatively unpopular, for there are various difficulties that obstruct the path to the smooth reduction and there are reasons to suspect that it cannot be carried through (for an overview, see Schiffer 1987: Chapter 9).

This list of main issues connected with the notion of content shows clearly that the content is supposed to accomplish very different things. Content is taken to be objective, exhibiting intentionality, it should be naturalised, and it is hoped that it plays a role in the explanation of action, mental processes and linguistic meaning. I am not sure whether a single thing could fulfil all these roles, but for the purposes of the present book, this does not really matter. I brought these various issues up in order to make it clear in what sense I am going to talk about content. Namely, there are some important distinctions to be made. There is a distinction between *the possession of content* and *content*. The latter comprises everything that is represented by the representation, while the former concerns what is involved in having the respective state with its content. I shall defend and elaborate interpretivism at the level of content-possession and not at the level of content. There is no defensible and reasonable interpretivist account available about the content of belief. But things are different when we ask what conditions one has to satisfy in order to have a mental state with content. Here, I submit, and this is the main message of the present book, an interpretivist perspective can be illuminating. In short, the question that I am going to discuss is the following: *what is it to have mental states with content?* This should not be confused with inquiries into the nature of the relation of representing (the intentionality issue) and into the nature of the relata of the representing-relation (issues concerning the nature of propositions and possible worlds, etc).[3]

3. There are also various puzzle cases concerning the specification of one's belief contents as given in belief reports (see e.g., Kripke 1979) as well as a small industry devoted to dissecting these issues. The ascription theory is not presented as a competing account of the semantics of belief ascription among others. The main problems there originate from the opacity of sentences specifying belief contents as well as from the fact that the truth-value of one's belief may alter without one's being aware of it. My question is rather more general: what conditions have to be fulfilled for one to have a belief (and other mental states) at all.

A similar distinction can be also drawn between the possession of a state and the nature of a state. For example, the question 'What is it to have a belief?' differs from the question 'What is a belief?'. These and some additional distinctions become clearer when we look at possible answers to various questions concerning mental states and their contents.

Let us consider possible stories that can be offered about S's believing that *p*. It is better to take believing as an example, since it is more specific than just being in a mental state. Imagine that the question is deliberately unspecific, so it would provoke as many different kinds of answers as possible. Taken individually they might not be suitable to complete the stricter frame of 'S believes that *p* if and only if…' Perhaps a combination of different answers would cover enough ground for this. However, at this point I want to cast my net as wide as possible in order to bring out distinct but still somewhat connected issues. We may thus simply ask what is involved in S's believing that *p*.

We get the following non-exhaustive list. In the first family, there are the following answers from (1) to (3):

1. S is committed to *p*
2. S has a high degree of confidence in *p*
3. S has the disposition to assent to *p*

These answers give the meaning of 'believing'. They centre on the state component of 'S believes that *p*', and are thus not rivals to the second type of answers:

4. S is in the believing-relation with the proposition *p*
5. S is in the believing-relation with the sentence 'p'
6. S excludes all non-*p* possible worlds

These are centred on explicating an element of the believing relation. In some accounts, (4) and (6) are not incompatible, for one can understand the proposition itself as a set of possible worlds. Although these answers deal with content, they do not touch on the issue of intentionality or mind-world connection. This is the topic of the third set of answers:

7. S's belief state token covaries nomologically with the instances of some physical property type *P*
8. S's belief state token fulfils the *P* type inferential role in S's cognitive system

These, in turn, are distinct from the fourth type, which provides a metaphysical story about what it is to believe that *p*.

9. S is in a brain state type *x*
10. S is in some token brain state
11. S is in a functional state of type *x*
12. S's believing that *p* is the most predictive interpretation of S's behaviour

A more apt question for the first family would be: what does 'believing' mean or what is believing? The second family is more appropriately viewed as a collection of answers to the question – how is content *p* represented? Here the issue is whether representing involves being related to a proposition, sentence, etc. The ascription theory is not concerned directly with this issue. The third family gives us various stories about what makes it the case that the belief that *p* represents *P*-s. And the fourth class properly concerns the question of what it is to have a belief that *p*. Although the question for the third family can in some cases be posed in the same way, it is crucial to distinguish the third and the fourth class. For only then does it become visible that the rivals of the ascription theory, which belongs to the fourth class, are not accounts of content like teleological theory, causal theory or functional-role theory, which make up answers of the third family. Indeed, if suitably interpreted, they can in principle be coupled with a kind of interpretivism.[4] These accounts state what determines the content of mental states, while the ascription theory states what it is to have mental states with content. Of course, when they make assumptions on the latter topic, they become rivals to the ascription theory. For example, a causal theory would require the existence of internal tokens standing in causal relations with the objects they represent. It is common to take these tokens as mental representations the possession of which is independent from any interpretation. In such a case, we have an instance of coupling theories of content of the third kind with more general assumptions about the possession of mental states (the fourth kind). The latter are the direct competitors to the ascription theory, which is neutral in respect of the three other types. The accounts of the fourth family all present an alternative answer to the possession-question. I discuss some of these approaches like functionalism and identity theories in the next chapter.[5]

---

4. For example, Davidson marries his interpretivism to a causal theory of content. Daniel Hutto (1999) holds the teleological theory of non-conceptual content, but combines it with a version of Davidsonian interpretivism. It is another question whether the result in each particular case is preferable on general grounds of coherence and simplicity.

5. Brandom (1994) gives an account of the meaning of 'believes' that belongs to the first family, and by this, he also indirectly answers the fourth question. I will provide an argument against such an approach in Chapter 5. Van Gelder (1998) proposes that we should refrain from looking for the type-4 account of belief, which is hampered by mistaken assumptions, and that the only good prospects in tackling belief are found in the analysis of its meaning, which results in answers of the first type. The difference between my approach and Van Gelder's might just be a difference in emphasis. Although the ascription theory provides the type-4 answer, it is far less metaphysical than common mind-body theories and it rejects the assumptions generally held by the type-4 theorists.

Given the distinction between the theory of content and the theory of content-possession, it should be evident that I am not going to provide a substantial account of intentionality. After all, the issue of intentionality was precisely the question of what makes it the case that mental states can be about the world. Why should we not give an interpretivist account of such a relation? Of course, one can do this, but it would not make much sense. Imagine that you employ a simple account of interpretivism, which states what makes it the case that one has a belief that *p* is that belief that *p* can be ascribed to one in manner *x*. If this is used to explicate the intentionality issue, then we get the following quite absurd result: my belief that *p* is about *P* if and only if belief that *p* can be ascribed to me in *x* manner. Since aboutness is closely related to correctness conditions, we can say in other words that on this account, the correctness condition of content is determined by the correctness condition of content ascription. In that case, my belief that snow is white is correct (or about white snow) just in case it is ascribed to me in *x* manner. My belief that the sky is blue is correct (or about blue sky) just in case it is ascribed to me in *x* manner. And so on, for any other mental state with content. This fails to individuate contents and it does not provide any illumination of aboutness.[6] Hence, the interpretivist approach to intentionality is not viable.

As noted, interpretivism can be coupled with different accounts of intentionality. In this book, I refrain from taking sides on the issue of which account of the mind-world connection is the right one. All I am going to assume is that content *p* is about *P*, leaving the substitution instances of *P* dependent on one's ontology. The belief that snow is white is about snow being white. There have been attempts to give a substantive explanation of the mind-world relation, but I am not convinced that any of them are completely successful. To argue for this would need a separate book. Presumably, this stance becomes more palatable if we draw a distinction between aboutness and meaning. The question of aboutness, which is also closely related to the issue of truth, asks what makes *p* about *P*? The question of meaning asks for the meaning of 'p'. I shall discuss these issues further in the next section. Here I would only like to note that while in my view there is a substantive story to be told about the meaning, the prospects of explaining how *p*-s could be about *P*-s seem rather bleak. This is not meant to suggest that there cannot be a theory of reference. Such a theory would state what is involved in fixing the reference and would specify referential conditions for different types of terms. What I object to is a more substantial and metaphysical view on aboutness. A metaphysical view on aboutness presumably rests on the picture of mind that consists of internal states, separated from the external world. Given this view, the

6. Davies (1995: 295–296) levels a similar charge to the Dennettian intentional stance approach, but he ties this with possession-conditions for concepts.

fundamental problem is how to connect those states again with their objects. I expect that this kind of problem would dissolve, once its presuppositions are laid bare and discarded. In addition, there is always some scientific story to be told. For example, those organisms that have highly specialised brains are also wired to detect certain properties or states in the environment. These property-detectors are specialised to certain properties. In this way, we can explain the mechanisms by which the brain reacts to certain stimuli. I do not think that there is much extra philosophical work to be done in this respect apart from trying to dispel certain metaphysical images that lead to unanswerable worries.

I leave the issue without final resolution here. What has to be pointed out, however, is that if the foregoing thoughts are on the right track, then a satisfactory account of content can be given by providing merely an account of content-possession. There will be no additional substantial task of explaining the nature of aboutness. But should this assumption be overthrown by some successful account of the mind-world relation, the status of interpretivism need not be shaken thereby. For provided that this account of content has no commitments that conflict with interpretivism, the latter can complement it as an account of content-possession.

What about the other aforementioned issues related to content? If the task is to provide an explanation of content-possession, are there any implications about objectivity and explanation? The task of explaining the possibility of communication and objective knowledge is not in the foreground here. There is some connection with the explanatory role of content though. Namely, the fact that we cite contents in explaining action is an important element of folk psychology and it cannot be ignored. I shall come to this in Chapter 7, where I try to find a place for causal explanation in the interpretivist framework. Matters are different with the explanation of linguistic meaning and communication, as well as with the postulation of the language of thought. I do not think that mental representations are explanatorily primary with respect to the states of common-sense psychology. On the contrary, the use of the notion of content in characterising mental representations derives from its use in common-sense psychology. The ascription theory, as I am going to present it, presumes that words have meaning, so it cannot be used in turn to explain the meaning on pain of circularity. This is the topic of the next section.

1.2 Truth and meaning

My aim in this book is to develop and defend a version of interpretivism about the possession of mental states and contents. It would be detrimental to this aim if this account were heavily dependent on some controversial account of truth

or meaning. Therefore, whenever possible, I have attempted to minimise the commitments of this account in the adjacent areas. For example, in Chapter 2, when I outline some metaphysical preliminaries for my study, I employ the notion of a natural basis, while leaving the issue of the ontology of its elements largely open. Particulars, powers, properties, etc can form the natural basis, and there is no need to make more specific commitments in the present context. The same applies to the notions of truth and reference. The main points of the ascription theory do not depend on a specific theory of truth or reference. However, I assume a broadly deflationary stance on truth that takes '*p*' and '*p* is true' as equivalent (there are different deflationary accounts, for a recent minimalist view, see Horwich 1998b). This stance manifests itself in the fact that I do not attempt to clarify meaning in terms of truth and I generally avoid phrasing issues in terms of the truth of some propositions. Note that if we presume this approach to truth, and if the deflationary scheme applies in the given discourse, there will be no substantial distinction between such discourses in terms of their truth-aptness. In particular, it cannot be said that the folk-psychological discourse is non-cognitive or that its sentences lack truth-values, for, if nothing else, they would be still apt for the deflationary truth.

Whereas my approach to truth is deflationary, I take seriously the notion of meaning and think that there is a serious task of working out a substantial theory of meaning. As concerns the ascription theory, however, all we need to assume is that terms already have some kind of meaning, but further details of the theory of meaning are not required. The presentation of the ascription theory does not presume that we have a fully worked out account of meaning at hand. I will return to this topic in Section 5.1.

The ascription theory is not proposed as an explication of the concept of having a belief or the meaning of the term 'having a belief'. Nor does it offer a theory of truth for statements about beliefs. Its aim is just to say what makes it the case that X believes that *p*. To see the difference, let us consider an example. Consider two claims:

(A) 'Mary has a little lamb' is true if and only if Mary has a little lamb
(B) Mary has a little lamb if and only if a little lamb has been acquired in some way by Mary, etc.

(A) gives us a deflationary truth condition for a sentence 'Mary has a little lamb'; (B) makes explicit a part of the condition that must be fulfilled for Mary to have a little lamb. Note also that (B) presumes that the term 'lamb' has a (prior) meaning. The ascription theory is similar to (B).

In the previous section, I noted that the distinction between content and content possession can be generalised into a distinction between explaining the

nature of an entity and explicating its possession condition.[7] In some respects, this distinction is quite trivial. Just compare the two questions: 'What is *X*?' and 'What is it to have *X*?' When we study the nature of an entity, the focus is on its constitution and individuating features. When we pose the question about the possession of some entity or property, then we are interested in those conditions that have to be fulfilled in order to make it the case that someone has this entity or property. In some cases, the distinction does not seem very helpful. For, it can be said, if we already have an account of the nature of an entity, then the account of the possession simply follows from the former – it is to have an entity with the specified nature. I do not attempt to argue that this distinction is useful across the board. However, the distinction becomes relevant when the entity under discussion is understood only in a deflationary manner (as outlined in Chapter 2). In that case, it has no nature other than what can be gleaned from the meaning of the respective term. Then, the account of the possession condition can also be viewed as an account of the condition that states when the respective term is applicable, and at least in the case of deflationary entities, this condition can be a substantial one.

I presume in this book that mental entities are entities in the deflationary sense. The ascription theory is presented as an account of what it is to *have* mental properties. Arguably, the possession condition is not part of the meaning of the term. For some such entities in the deflationary sense, facts concerning their *possession* may involve recognition-dependence, whereas the specification of their *nature* does not make them recognition-dependent in this way. For example, let us assume that 'desiring *p* to occur' means 'wishing strongly that *p* will happen', which uncovers the recognition-independent nature of this state. An interpretivist account of what it is to have such desire, however, makes the possession of this state recognition-dependent, given that it makes the possession dependent on an interpretation.

Since we can grasp the meaning of mental terms without having a clue about the ascription theory, the central claims of the ascription theory must be something additional to the meaning. In my view, the meanings of mental terms are fixed in the language independently from the ascription theory. The latter plays a central role in explaining what is involved in the possession of mental entities and the main idea is that the subject must be describable in mental terms (no matter how their meaning is fixed) or, what amounts to the same thing, that these mental entities must be ascribable to the subject.

---

7. My notion of the possession condition and the way I use it differs from Peacocke's (1992) notion, who individuates concepts in terms of their possession conditions.

At this juncture, some comments on my use of the term 'concept' are in order. For the most part, I use 'concepts' in the non-technical sense as 'notions' or 'public meanings of terms' and not in the sense of internal mental representations or other individual modes of presentation. Neither do I assume that concepts are the same as universals. Thus, when I talk, e.g, about the concept of belief, then I have in mind the notion of belief as it is widespread in our linguistic community and not the idiosyncratic conceptions one may have about beliefs. The individual aspects come into play when I discuss concept-possession. However, I do not think of the possession of a concept as a fact about someone that is itself independent from the ascription. My view, which I touch upon in Chapter 8, but do not develop fully in this book, is that the possession of concepts is derivative with respect to a longer story that can be presented about one's linguistic and behavioural dispositions and that it does not explain the latter.

1.3 A road map

Since this is not a detective story, where the culprit becomes known on the last page, I finish this introduction with an overview of what will happen in the coming pages. The book is divided into three parts. The division follows the structure of the book, so that people with different interests could begin by reading it from different parts. The first part includes Chapters 2 and 3. It provides mainly a review and a criticism of existing philosophical positions. Thus, those readers who need to be convinced that interpretivism is a position worth exploring could carry on reading Chapter 2, which criticises the main alternative views. Those who are interested in interpretivism could go straight to Chapter 3. Those readers keen to explore my more specific version of interpretivism should continue reading the book from Part II. I would advise the latter readers also to take a look at Section 2.1, as this presents a metaphysical framework that will be presumed in the following discussions.

Now for a more detailed overview of the chapters. In Chapter 2, I criticise the alternative approaches to the question of what it is to have a mind. I begin by presenting a general metaphysical scheme for the following discussion, which rests on the distinction between pleonastic entities and the natural basis. I hope that this distinction gives us a better grasp of various positions on the mind-body relation, and allows us to highlight the point that we do not really need to introduce (and have no justification for introducing) mental properties in any stronger sense than the pleonastic. But if that is so, the mental does not provide an ontological problem. All that has to be explained is the status of pleonastic mental properties. These metaphysical preliminaries are followed by an analysis and critique

of several prominent accounts of mind-body relations: supervenience, identity theories and versions of functionalism.

Chapter 3 is a first step in my exploration of interpretivism – the view that interpretation is involved in determining what mental states and contents one has. I provide a thorough overview of two major interpretivist approaches to mind and content put forth respectively by Dennett and Davidson, to which the ascription theory is indebted. However, I also present some objections to Dennett and Davidson and show how my approach diverges from their accounts. I find it more straightforward to elaborate the ascription theory as a largely self-standing position, not as a piecemeal development of Dennett's or Davidson's proposals. For this would add an additional constraint of keeping my claims consistent with some of their general presumptions that I might not find plausible. Instead, I have proceeded only from what I deem tenable and worth defending. Some of the claims I present and defend here have been also urged by others. What is new here, I hope, is the ways these ideas have been put to work in defending interpretivism and the details of the general framework that emerges in the pages that follow.

Sometimes, interpretivism is presented as an example of an antirealist approach to the mind. My contention is that the general label of realism or antirealism is not particularly helpful in the present setting, since it can be used in a variety of senses and often such discussions remain too vague and general without further specification. I tackle some aspects that have been connected with realism like recognition-independence and objectivity of mental patterns in various chapters, especially in 3 and 5, where I try to the give these issues more specific content.

In Part II, over the course of the next three chapters, I present my own version of interpretivism. Chapter 4, "Folk psychology and mental terms", does some preparatory work. I elaborate the sense in which we can understand mental states as folk-psychological states, and argue against some alternative views. Chapter 5 is in many ways the central chapter of the book, for it outlines the ascription theory, which is a basis for the discussion in later chapters. The theory explains the possession of mental states and their contents in terms of their canonical ascribability on the basis of evidence from various sources. An ascription is canonical when it conforms to specific requirements concerning coherence and revisability, as outlined in that chapter. Chapter 6 presents several potential objections to interpretivism and the ascription theory and answers them from the ascriptionist perspective. Finally, I discuss the ways of arguing for the ascription theory and interpretivism in general.

The last three chapters comprising Part III concern applications of the ascriptionist account. When applying the account to different topics, my purpose is to demonstrate its viability by showing that it can provide solutions to a variety

of burning issues in the philosophy of mind. Chapter 7 outlines the problem of mental causation. First, I explain why mental causation poses a problem at all and then proceed to work out an account of causation which would not compromise the causal standing of mental states and which would be consistent with the interpretivist approach. In Chapter 8, I discuss the application of the ascription theory to perceptual states and contents. The aim is to show that the theory need not be confined only to standard propositional attitudes such as beliefs and desires. Chapter 9 presents an account of self-knowledge that accords with the ascription theory and satisfies our pre-theoretical intuitions of self-knowledge. In that chapter, the main challenge is to show that contrary to the widespread assumption, interpretivist accounts are compatible with both first-person authority and a subpersonal explanation of self-knowledge.

PART I

Towards interpretivism

CHAPTER 2

How not to have a mind

Mainstream philosophy of mind is replete with attempts to explain mental properties in terms of other, naturalistically acceptable kinds of properties. This chapter attempts to make sense of various relations between mental and physical (including functional) properties. The aim of this discussion is to outline more clearly some potential alternatives to interpretivism as well as to present arguments against them. By showing that other positions in the metaphysics of mind leave much to be desired, I strengthen the case for interpretivism. In addition, this helps to place my own views on the map in a more straightforward manner in the chapters that follow. Before we can proceed to study mind-body relations, however, there is one more task to be undertaken. Namely, it is necessary to work out a general framework for discussing positions in the metaphysics of mind. This framework will be assumed throughout this book; indeed, several of my discussions in later chapters rely on it. To this aim the next section is devoted.

2.1 A general metaphysical scheme: Pleonastic properties and the natural basis

When the nature of mental states is discussed in contemporary philosophy of mind, the discussion is usually ambiguous in the metaphysical sense. The discussion is sometimes framed in terms of mental states or properties and centres on their relation to physical states or properties, but occasionally similar points are made in terms of predicates. It is the nature of this relation that has captured most of the attention rather than the nature of relata, that is, notions like 'predicate' and 'property'. The common assumption is that the same issues would arise whether we talk about mental predicates, properties or states. It is true that similar questions would arise, but there is no guarantee that the answers would be similar in each case. The reason is that the relations between properties need not mirror the relations between predicates. For instance, there can be a relation of metaphysical dependence between property-kinds without any conceptual relations between the predicate-kinds by which we pick out these properties.

A characteristic example of the readiness to switch from predicates to properties is evident in the following passage by Jaegwon Kim:

> For Davidson, the content of the claim that all mental events are physical events turns out to be only this: every event that can be given a mental *description* can also be given a physical *description*, or, *as we might say in current idiom*, that every event that has a mental *property* (or falls under a mental kind) also has a physical *property* (falls under a physical kind). (Kim 1998: 4, my italics)[1]

I do not object to this idiom of using the notion of property, and set up a scheme in this chapter that allows us to speak of mental properties in this sense. However, it is of crucial importance that one should not illicitly shift from this usage to the stronger notion of property, according to which it is not the case that for every predicate there is a corresponding property. My claim in this chapter is that if one distinguishes different senses of 'property', then this helps to distinguish real issues from verbal puzzles and makes it easier to assess various answers to the mind-body problem.[2]

The distinction between two senses of 'property' is not new (see e.g., Schiffer 1987).[3] We may understand properties in the deflationary sense and in the inflationary sense. Colloquially, speaking of *x*'s properties in the deflationary sense is just another way of talking about what can be predicated of *x*. No further assumption that an object has the property in some stronger sense, as something corresponding to the predicate, is required. In the inflationary sense, saying that *x* has some property conveys something about the way *x* is, and this is not dependent on the terms we use to specify this way.[4] Furthermore, in this usage, it is commonplace to assume that the applications of nonequivalent predicates converge on the same inflationary property. For example, we use such distinct predicates as 'being water' and 'being H_2O' to pick out the same property (assuming that water does not contain other elements). This distinction will be further clarified in the following pages.

1. It should be mentioned, however, that later in his book, Kim (1998: 105–106) comes to exploit the relevance of the property-predicate distinction.

2. I am talking mostly about properties, but it is assumed that similar points apply, *mutatis mutandis*, also to mental facts, events and states.

3. Lewis (1983a: 346) also distinguishes between two kinds of properties. However, his distinction between sparse natural properties (universals) and abundant properties (classes) does not coincide exactly with my distinction.

4. In one sense, the distinction between inflationary and deflationary properties might be confusing, since there is a profusion of deflationary properties. But here and also in the case of truth, the distinction is made in terms of metaphysical "weight". Deflationary properties are metaphysically insubstantial, but when inflated, they become substantial, inflationary properties.

Hijacking a term from Stephen Schiffer (2003), I shall use the expression '*pleonastic* property' when I talk about properties in the deflationary sense, which will be articulated more fully below. The word 'pleonastic' means something excessive or superfluous. Indeed, pleonastic properties are superfluous in the sense that talk about things having pleonastic properties can be replaced by talk about predication. On Schiffer's usage, all properties are pleonastic; however, I shall also occasionally use the term '*natural* property' to denote a subset of inflationary properties. In fact, terms like 'natural basis' or 'natural existent' seem more appropriate, for they do not commit one to any specific views about the metaphysical structure of this basis in a given case – it may consist of universals and/or particulars and/or powers and so on. Hence, when I use 'natural property', I leave it largely open how exactly natural properties are to be analysed and, mostly, I talk about the *natural basis* of a particular ascription or predication and entities that belong to this basis. For example, one need not presume that natural properties are abstract objects, although one can hold such a view in the present scheme. However, by adding the modifier 'natural', I suggest that one can make sense of properties in a broad naturalist picture, thus barring pure platonic universals. When I talk about 'using the notion of property in the inflationary sense', I want to stress the usage of the notion of property, which aims to pick out some existing entity and also to highlight that this may not be justified for this specific kind of properties. I shall say more on this in what follows.

Schiffer's (2003: 50–71) account gives a nice explication of the deflationary notion of a property. In this regard, it is reasonable to give a short outline of his treatment here, but the more specific details of his proposal are not essential to the points that I am making in this chapter. For my purposes, it is sufficient to keep in mind that for every well-formed predicate, there is a pleonastic property.[5]

According to Schiffer (2003: 57), "a *pleonastic entity* is an entity that falls under a *pleonastic concept*". In his view, properties, facts, propositions, events and fictional entities are all pleonastic entities. A concept of an *F* is pleonastic if it implies "something-from-nothing" transformations, as outlined below, and adding this concept and its transformation-claim to a theory results in a conservative extension of that theory, given that the quantifiers of the theory are restricted

5. The qualification 'well-formed' is added for pragmatic purposes to rule out negative predications and bizarre conjunctive predicates. However, since pleonastic properties carry no metaphysical weight, there is no principled reason why one should not have a pleonastic property of being not a giraffe. Insofar as such predications are not common practice, there is no need to worry about such pleonastic properties either.

to non-*F* entities.[6] One theory *T´* is a conservative extension of another theory *T* in case "*T´* includes *T* and nothing statable in the vocabulary of *T* is logically entailed by *T´* but not by *T*" (Schiffer 2003: 54). If there are statements in the new theory that can be also stated in the vocabulary of the old theory, but were not entailed by it, then the new theory is not a conservative extension.

What is a something-from-nothing transformation? This is a case of inference, where a statement containing a singular term is inferred from a statement that does not contain that singular term. For instance, to consider an example based on Schiffer (2003: 61), we take a statement with a singular term 'Stephen'

(1) Stephen is clever

and infer from (1) another sentence containing two singular terms 'Stephen' and 'the property of being clever'

(2) Stephen has the property of being clever.

By this transformation, we have introduced a pleonastic property of being clever. According to Schiffer, this is a valid inference, since he takes it to be a conceptual truth that if S is clever, then S has a property of being clever. The reason why this is a conceptual truth, according to him, is that such transformations constitute our conceptual practices with the concept of a property. Hence, it can be said that if one possesses the concept of a property, then one would also regard any such transformation as valid, irrespective of its singular terms.[7]

Schiffer's scheme demonstrates how easy it is for us to move from talk that does not mention properties to property talk. This easiness shows that earning the right to talk about properties in such a manner must be insignificant to the enterprise of determining a theory's ontology. In Schiffer's (2003: 63) words, adding pleonastic entities to a theory "does nothing to disturb the output of that theory, does nothing to alter that theory's take on the pre-existing causal order". Now, Schiffer stresses that the something-from-nothing transformation creates new referential terms. Although this is central in the paradigm of determining the ontology of a theory by checking the reference of its statements, this particular machinery is not essential to the notion of a pleonastic property as I understand

6. This restriction is needed to block inconsistencies that arise when the theory contains certain claims about F-type entities (see Schiffer 2003: 55–56). The "conservative extension" strategy originates from Field (1980).

7. Schiffer's something-from-nothing transformation for pleonastic facts is similar. It allows us "to move back and forth between 'Michele is funny' and 'It's a fact that Michele is funny'" (Schiffer 1987: 51). But it has to be noted that there are some differences in the treatment of pleonastic entities between Schiffer (2003) and Schiffer (1987).

it here. Moreover, one can argue that by applying the criteria for ontological commitments, one cannot determine which natural properties there are, and hence cannot fix real ontology (e.g., Macdonald 2005: 29). I shall say more on this later in this section, where I argue for the need for substantive ontology and why we should not confine ourselves solely to the pleonastic realm. But for now, the lesson is that there are pleonastic properties in virtue of the fact that we can rephrase statements involving predication into statements ascribing properties. For this idea, Schiffer's points about something-from-nothing transformation as well as about conceptual truths, putting his more specific pronouncements aside, give good expression and motivation.

We can thus say that whenever a predicate applies to an object, an object has a pleonastic property. In this sense, saying that an object has a property and saying that a predicate applies to an object are interchangeable. Saying that an object has a property does not carry any metaphysical burden in such a usage – nothing more than that a predicate applies to this object follows from this.[8] This point about the lack of metaphysical consequences of deflationary properties becomes fairly central in some of the following sections. Of course, there is a story to be told about the application conditions of the predicate, and there is no reason to assume that this story consists just in the fact that the object has this property, for that would be uninformative, if not circular.

Such a notion of a pleonastic property follows the principle that is made explicit and severely criticised by John Heil (2002, 2003a, 2003b). It is the following principle Φ:

> (Φ) When a predicate applies truly to an object, it does so in virtue of designating a property possessed by that object and by every object to which the predicate truly applies (or would apply). (Heil 2003a: 26)

According to Heil, this is a deeply entrenched but unfounded assumption in analytic philosophy, and it is responsible for conceptions like higher-order properties, supervenience and uncritical multiplication of levels of reality. He also thinks that once we drop the principle Φ, and replace it with a weaker claim that a predicate applies to an object in virtue of *some* of its properties, we would also dissolve many of the problems with the aforementioned conceptions.

---

8. Cf. L. R. Baker's (1989: 305) usage of 'feature' in analysing Dennett's intentional stance: "I shall use the term 'feature' with next to no ontological commitment: a system *S* has a feature *F* if and only if sentences of the form '*S* is *F*' are true." It is interesting to note that Davidson (2001a: 299) also resorts to a non-committal kind of speaking: "Having banished properties, states, beliefs, and so on, from my explanatory machinery, I shall now continue to allow myself to use words that superficially seem to refer to such things, on condition that I be understood to hold that such talk can be exchanged at boring length for talk that doesn't use these words as referring to entities."

I agree with Heil that there need not be any one-to-one correspondence between predicates and properties, taken in the *inflationary sense*. Indeed, this is where the principle Φ does its damage, as pointed out by Heil. For if one takes predication in general to be a reliable guide to ontology and implicitly holds onto the principle Φ, then one ends up with a myriad of inflationary properties and is pressed hard to make sense on their inter-relations. There is a standard naturalist assumption that only properties at lower layers are ontologically kosher and in order to accommodate higher-order properties, those have to be reduced or otherwise related to more basic properties. This principle and such reasoning have repercussions for all main theories in the metaphysics of mind. Consider an example adapted from Heil (2003a: 7). Once we have assumed the existence of the inflationary property of being in pain, and are unable to find a corresponding physical property, the principle Φ leaves us only two choices: either it is not true to say that someone is in pain and there is no such property, or being in pain is a higher-order property. But the notion of an inflationary higher-order property is deeply problematic (see the discussion of role functionalism in Section 2.5). On the other hand, if we understand the principle Φ as governing the use of the notion of a pleonastic property, and keep in mind that talk in terms of pleonastic properties carries no immediate ontological commitments, the principle Φ does no harm.

Then again, one need not confine one's claims only to pleonastic properties, for the truth of statements can be grounded in various ways. Depending on the nature of the statement, it can be grounded on conceptual relations between concepts, or on facts concerning the application conditions of concepts, or on natural relations between entities in the natural basis, or on the relation between pleonastic properties and the natural basis. This presumes the distinction between pleonastic entities and their natural basis, but it does not entail that pleonastic entities are legitimate only when they can be directly derived from or related to the natural basis.

Reasons for drawing the distinction between pleonastic properties and the natural basis are actually widespread in the literature. Basically, these are the same arguments that are developed against the strong form of predicate nominalism, which would be quite an apt name for an account of pleonastic properties, if this purported to give the whole story about properties. This is the view that there is nothing more to the nature of a property than the applicability of a predicate. Predicate nominalism is sometimes formulated (see e.g., Binderup 2004: 63) as the following reductive biconditional (i.e., the right-hand side gives a complete reductive analysis of the left-hand side):

x has the property *P* iff the predicate 'P' applies to *x*

A common objection to this biconditional is that it leads to object regress and relation regress (Armstrong 1978: 18–21). The first is the problem that the nominalist

needs some ground to talk about predicates of type *P*, which covertly brings in again some properties that are shared by the tokens of this type. The second is the problem of explaining in nominalist terms the relation of predicate-applying-to-an-object. This also constitutes a type that is shared by all the objects to which 'P' applies. Of course, the nominalist might distinguish between properties and types, and argue that the existence of types does not entail the existence of unanalysed properties and explicate falling into the same type as a membership in a resemblance class (which in turn has to be explicated in a non-circular way). But this move is not open to the predicate nominalist of the simple variety. For, given her claim that there is a property whenever a predicate applies to an object, she cannot deny that there is an unanalysed second-order property, for there is a second-order predicate that applies to objects of the same type, namely, the predicate of predicate-'P'-applying-to-it.

These problems show that a thoroughgoing predicate nominalism is not a stable position. I shall not discuss this objection further, since I agree that one needs to have some construct in one's ontology that allows one to make sense of types and repeatables. Unlike Schiffer, I do not think that pleonasticity exhausts the nature of properties (or their ontological siblings). I can thus explain that in general, objects with the same or similar pleonastic properties belong to the same type in virtue of sharing some of their natural properties, and since types themselves can be viewed as pleonastic descriptions, I am not obliged to treat types as natural existents or universals.[9]

It is also objected to predicate nominalism that the applicability of a predicate neither implies nor is implied by the existence of a corresponding property. In other words, it is thought not to be the case that whenever an object has a property *P*, there is a corresponding predicate 'P' applicable to this object and not to be the case that whenever the predicate 'P' applies to an object, it has a property *P*.

As concerns the latter, the common intuition is that there are various predicates for which there exist no corresponding properties. I have touched on this problem already – given that properties are pleonastic, there is no principled reason to object to 'weird' properties. But, if one insists, then the scope of the biconditional can be restricted to some set of well-formed predicates. The other horn – that properties do not imply predicates – is fuelled by the intuition that there can be properties for which we lack predicates. It does not make sense to

---

9. This is more complicated in the case of mental types. In my view, two subjects can both have a pleonastic mental property of the same type (e.g., believing that snow is white), while sharing no relevant natural properties in their bodies or brains. The matter of belonging to the same type is thus solely a matter of being describable with the same mental concept or predicate, and this is determined by a complex net of social and conceptual relations.

suppose that for every way the world could be, there is a concept available to express that mode (cf. Heil 2003a: 72). Again, given the manner in which pleonastic properties were introduced, there is no reason to share this intuition. It is trivially true that the set of *pleonastic* properties is limited to the distinctions that can be made in a language.[10] However, this answer is available to me only for the reason that I do not claim that pleonastic properties exhaust all there is to say about the way the world is. What is really behind the intuitions concerning the lack of mutual implication is the idea that language does not determine the ways of the world. Hence, those who share this intuition, including me, have reason to claim that not all properties are pleonastic, or at least that there is a need for a construct that enables us to point to a world that is independent from our conceptualisations.[11] Therefore it is important to note that what is usually said about the mismatch between predicates and properties applies in my picture rather to the relation between pleonastic properties and natural properties. I will return to this issue in a few paragraphs.

Consequently, the pleonastic account cannot be the whole story about properties. It may well be that something-from-nothing transformations govern some part of our use of the concept of a property, but there is another aspect that should not be overlooked. This is the fact that by invoking the notion of a property, we aim to point out how things are in the world. That is something different from saying merely that a predicate applies to an object. One is prone to overlook this aspect if one has nominalist inclinations and an undifferentiated notion of a property. Given the assumption that there is a property for every predicate, there seems no need to go on and make additional claims about properties that the predicate 'really' picks out. But once we distinguish between deflationary and inflationary properties, and keep the principle Φ for the former class, but drop it for the latter class, then there arises a need to say more about the latter, the natural basis, as I call it. We need to make room for more substantive ontology. Of course, the final

10. Schiffer takes pleonastic entities to be language-created but language-independent. To explain what this means, he imagines a possible world with people without the concept of a property. He claims that "these people are entirely ignorant of properties, even though they live in a world as rich in properties as the actual world" (Schiffer 2003: 62). By contrast, I think that pleonastic entities are dependent on the existence of language in the following way. We are entitled to describe the possible world "as rich in properties" only because in the actual world we have fixed the language-schema to describe the situation elsewhere and in this schema, we have the concept of a property. Had there been no language-using beings in any possible world, there would have been no pleonastic entities, but, of course, there would have been the natural basis.

11. This claim does not conflict with the idea that there can be different basic ontologies, for once one adopts an ontological scheme, one can say inside that frame of reference that it captures the real categories.

ontology does not have to make use of the notion of a property, if like Schiffer, one wants to reserve that term solely for pleonastic properties, and if one assumes that 'property' implies something akin to platonic universals and is suspicious of the latter. However, some related categories like those of powers or something similar are almost certainly needed to state its central principles and I see no good reason not to call these 'natural properties'.

Since Schiffer takes all properties as pleonastic and all facts about properties as constituted by our practices constituting corresponding concepts, he has argued himself to the point where there can be no determinate answers to most metaphysical questions. Is the property of being a dog identical to the property of having such-and-such a genotype? For Schiffer, there is no determinate answer, since our conceptual practice does not support such identity-claims. He follows the general principle that "if a question of individuation is left unsettled by the practices constitutive of the concept of a property, then that question has no determinate answer" (Schiffer 2003: 63). But this just has the general consequence that there are no determinate answers in metaphysics, for the very fact of philosophical disagreement shows that our conceptual practices do not settle matters. Moreover, the issue of property individuation is ambiguous between the individuation of pleonastic and natural properties. Pleonastic properties are individuated, having concepts in view, and they inherit their conditions of individuation. Note that an account of concept individuation should allow that synonymous predicates would give us the same pleonastic property.[12] But the individuation of natural properties is not constrained by linguistic facts: it goes by what there is, perhaps by causal powers. In my scheme, we can still make sense of the possibility of metaphysics. The property of being a dog and the property of having such-and-such a genotype come out as distinct pleonastic properties if the corresponding concepts are distinct. Their being distinct is compatible with the fact that these two pleonastic properties have the same natural basis, i.e., they correspond to the same state of affairs or to the same arrangement of powers in the world. There are thus two cases of individuation. Only in those cases where the existence of the natural basis has no bearing on the possession of a pleonastic property should we individuate properties by relying solely on the nature of practice in order to determine the sameness of respective concepts.

12. Cf. Jackson (1981: 295), who held that properties are identical, if the predicates that are used to ascribe these properties are synonymous. He maintains this claim only for *property-ascribing predicates*, and hence is not committed to saying that there is a property for every predicate. I agree with this, when it is applied to properties in the deflationary sense, and disagree, when it is thought to give us inflationary properties.

The exact details of the natural basis do not matter in the context of the present chapter. However, I make clear some of my assumptions in this regard. By drawing a distinction between pleonastic and natural properties, I have joined those philosophers who reject semantics as a direct guide to ontology. Semantic analysis itself, that is, making explicit whether our terms function as referential expressions; finding out the ontological commitments of statements etc., does not bite deeper than the deflationary layer.[13] In this respect, the explication of ontological commitments of theoretical statements might not cut deep enough. Ontology is best taken as a constructive discipline aligned with the reflection of our finest scientific theories.[14] A compelling example of work in this direction is the scientific essentialism of Brian Ellis (2001, 2002); other prominent examples are Armstrong (1997) and Heil (2003). In general, this understanding of ontology is in the spirit, if not the letter, of those people who investigate truth-makers, along with the rejection of the principle Φ for natural entities. For natural entities are rather more sparse than abundant. By such study, we can single out various candidates for natural bases. It is important to allow that the natural basis itself need not be a single layer matter. I would leave it open whether the natural basis consists solely of natural kinds, but it is true that what are commonly taken to be natural kinds surely belong to the natural basis. And there are good reasons to hold that natural kinds themselves form hierarchies, so that the higher-layer natural kinds have powers different from the lower-layer entities and are consequently governed by different laws (see Ellis 2002 for a view that natural kinds have essential properties, which include powers).

Of course, in order to describe elements of the natural base, natural existents, we need to use some terms, and this would trivially give us pleonastic properties.

---

13. I do not intend to make any commitments in the theory of reference by the introduction of the notion of a pleonastic property. For example, the claim that there is a pleonastic property of being water does not imply that the predicate 'is water' necessarily *picks out* or *refers to* the pleonastic property of being water, as opposed to the pleonastic property of being H_2O, or as opposed to some further natural basis of water/H_2O. I do not claim that the predicates refer to pleonastic properties. The basic idea is just that by employing predicates, we earn the right to talk about properties, but it still remains open which natural basis those pleonastic properties have, if any. It does not follow from this that those terms that are usually taken as co-referential turn out to be referring to different pleonastic properties. Indeed, this basic idea can be further developed into a kind of deflationary theory of reference, as opposed to a substantial or metaphysical account, but that would be an extra step.

14. It is possible to reconcile these two approaches by viewing the process of constructing or exposing an ontology as a study of referential commitments of the best scientific theories. However, it would then depend on one's account of reference whether this would give us the basic building-blocks of the world.

So why cannot we just set up the issues as relations between kinds of pleonastic properties or between theories involving the respective concepts? I would agree with such a set-up, given that we would also add that not all relations between pleonastic properties hold in virtue of conceptual connection; some relations are grounded in the world. Without this addendum, we would be cut off from the world and end up with some form of conceptual idealism, which says that everything is a text or that predications make up the inflationary ways the world is.

Let us return to the point concerning the relation between pleonastic properties and natural properties. Since in my usage, pleonastic properties are counterparts of predicates, the previously mentioned dissociation of predicates and properties generalises to the dissociation of pleonastic and natural properties. The existence of a pleonastic property does not imply nor is implied by the existence of a natural property. Let us address first why pleonastic properties do not imply natural properties. There is no reason to assume that for every pleonastic property there is a corresponding natural property in the world. We can describe someone as being happy, whereas it could be that there are no natural entities in our ontology that would fulfil the role played by the pleonastic property of being happy. Similarly, if an object has two distinct pleonastic properties, there need not be two distinct natural properties in the object. For example, it can be said about me that I am over six feet tall and over five feet tall. Although I thus have two different pleonastic properties, there is a single natural basis consisting of my physical structure amounting to the length of 6'3". From this incongruity it does not follow that we should say that "strictly speaking", there is no such thing as happiness or that it is not true that I am both over six feet tall and over five feet tall. Any such error-theory stems from some misguided views on truth and predication.[15] Similarly, even if there is no natural counterpart of happiness, there is a substantial story to be told about the application conditions of 'being happy'.

Neither does the existence of some natural property imply the existence of a pleonastic property, for reasons analogous to the predicate-property case. If we assume that this basis need not necessarily be properly describable in ways available to any present-day language, then it is not possible to draw a something-from-nothing inference to introduce pleonastic properties. There should also remain some room for ontological relativity – namely, given that there may be no uniquely privileged way of describing the elements of the natural basis, there may thus be no unique set of pleonastic properties corresponding to a given type of a

15. Some good advice from Ellis (2002: 45): "Never try to argue from the fact that something is true of something that it denotes a property of it." There is no direct route from a pleonastic property to the natural basis.

natural basis. I do not think that this possibility makes the notion of the natural basis itself incoherent, as some might suppose.

Some readers might be puzzled as to why I do not employ the traditional predicate-property distinction. What is the motive for my talk about properties in two senses? Indeed, I would also prefer to go along with the more common usage, but there are principled reasons why my usage is to be preferred in the present context. First, and foremost, a distinction similar to the one between predicates and properties is needed in order to make sense of the contemporary metaphysics of mind. However, this distinction is quite often shadowed by the implicit and unsubstantiated adherence to the principle Φ. The result is the abundance of inflationary properties as well as an ambiguity between conceptual and ontological solutions to the mind-body problem. When mental properties are inflated, then a physicalist is compelled to find a place for them in the natural order. This is why reductionism is so hotly debated, for the implicit assumption is that if mental properties somehow fail to be reduced, then physicalism is in trouble. Note that no such conclusion follows if mental properties are understood in the deflationary way. Inflating mental properties is nothing but an easy way to assume a heavy metaphysical burden. Moreover, most of the proposed reductive solutions fail to attend properly to the troubles with inflationary mental properties, since those solutions in fact do not presume the existence of inflationary mental properties, for – as we will see later in the chapter – they can be formulated as requiring merely pleonastic mental properties.

I am not claiming that contemporary philosophers of mind fail to distinguish predicates and properties. However, some might think that this distinction is not even needed. One might happily concede that it is not the case that there is a property for every predicate, but the principle Φ is still valid at least for most mental and physical predicates, hence it does not matter whether we put the issue in terms of predicates or properties. However, my worry is precisely that we are not entitled to shift from mental predicates to the existence of *inflationary* mental properties.[16] As soon as the notion of property (and other such notions) is bifurcated, it becomes apparent that the existence of inflationary mental entities is not something one can simply presume as a starting point in discussing the mind-body relation. The burden of proof lies on those who feel the need for inflationary mental properties. As I hope to show in the course of this chapter, inflationary mental entities are not even required by most accounts of this relation. These views are plausibly formulated without commitment to inflationary

16. This applies to other mental entities as well, including mental facts, events, states, etc. Without supporting arguments, one is not entitled to move from true claims about mental features to the existence of inflationary mental facts, etc.

mental entities. I think that once this is made apparent, the idea that there must be inflationary mental properties would lose its impetus. If one becomes aware of the option of working with a metaphysically unproblematic notion of a pleonastic mental entity, one would see that by refusing to accept the inflation of mental entities, one can obliterate the worry of finding a place for inflationary mental properties in the natural world.

Alternatively, it can be proposed that we should frame the study of mental-physical relations either solely in terms of predicates or solely in terms of properties, but not both in the same breath. The idea is that we are faced with similar problems in both cases, so that when we ask how mental predicates reduce to physical predicates, we may ask, *mutatis mutandis*, the same question about properties. And it is assumed that there is no additional interesting issue between the level or stratum of predicates and the stratum of properties.[17] However, there is no reason to assume that we are dealing with the same notion of reduction in both cases, and that this is not the case will become evident in the following sections. In making sense of these notions, it becomes clear that the distinction between predicates and properties or its analogue is essential.

Now, if we pair the traditional predicate-property distinction with my denial of the existence of inflationary mental properties, I would be regarded as an eliminativist (which I am not), and I would be compelled to present all my claims in terms of mental predicates. This brings us to the second reason for my talk about properties in two senses. If one takes contemporary mind-body theories solely as analyses of mental and physical concepts, then one risks misunderstanding those accounts. (An example of this is Ludwig 2003: 25, who assumes that there is a property for every predicate and ends up claiming that according to the identity theories and the functional accounts, mental concepts are physical concepts.) Similarly, the ascription theory is not an analysis of concepts like 'mental state', 'belief', etc. On the contrary, it does not say anything novel about the meaning of such terms. The ascription theory takes them to mean whatever they mean in that part of the language which is systematised in folk psychology. If the point of ascription theory has to be framed in terms of predicates, then it concerns rather the application conditions of mental predicates than their meanings. However, if

17. I shall use the term 'stratum' in the context of distinguishing the level of the world (natural basis) and the level of pleonastic properties. This is needed in order to separate this distinction from more common micro-macro as well as personal-subpersonal distinctions, for which the word 'level' is preserved. In this usage, levels and strata cross-cut each other. For example, one's discussion of personal and subpersonal level entities might be confined only to the pleonastic strata. Similarly, natural properties can form hierarchies of levels, or layers as I would prefer to use a distinct term there as well.

I were to phrase my points in terms of the application conditions of mental predicates, I would again risk losing contact with most contemporary accounts, for these are mainly framed in terms of property-talk. Therefore, I think it is best to approach the issues in terms of pleonastic properties rather than predicates, and to leave open the possibility of incorporating natural properties to the analysis of rival accounts. Only with the differentiated notion of a property can one provide a sharpened analysis, which shows how these accounts can be intelligible at all.

2.2 Mind-body relations

Mind-body relations can be presented in a way that leaves the deceptive impression that they are relations between two kinds of properties – mental and physical – that involve only one stratum; and the issue is just to establish how these two kinds (or tokens of kinds) are related. This impression is reinforced by such claims as 'mental properties are identical to physical properties', 'mental properties supervene on physical properties', 'mental properties are realised by physical properties', and so forth. I argue in this chapter that in order to make sense of these relations, we need to distinguish the strata of pleonastic and natural properties. I shall provide reformulations of the main mind-body relations within my metaphysical scheme and present arguments against them. Of course, my arguments are not dependent on these reformulations; the latter are needed to clarify the commitments of different views. The distinction between pleonastic and natural properties also enables us to show that from this perspective, some so-called antireductionist positions on the mind-body relation are still variants of reductionism. I use 'antireductionism' here in the loose sense shared by contemporary philosophers of mind that mental properties are both conceptually and metaphysically distinct from physical properties, but depend on the latter. Of course, there are more specific formulations of reduction and reductionism. These will be explicated in Section 2.4.

In general, we study the following claim: 'mental properties are in R with physical properties' where R stands for relations of various sorts. In some cases, it is correct to take it as solely expressing the relation between pleonastic mental properties and pleonastic physical properties. This is the case where one set of predicates is conceptually analysed into another set. For example, this is the position of logical behaviourism, which provided either behavioural translations or definitions of mental concepts. In that case, the analysis was supposed to show that mental predicates are related to behavioural predicates, and thus one can say that there holds a conceptual relation (that of definability or translatability) between pleonastic mental properties and pleonastic behavioural properties. (Note

that this analysis does not imply anything about the existence of inflationary behavioural properties.) However, for familiar reasons, no one seriously supposes anymore that this relation holds: namely, behavioural properties (including dispositions to behave) are by themselves neither sufficient nor necessary for ascriptions of mental properties. Given the holism of the mental, no mental property can be reduced one-by-one to behaviour or behavioural dispositions, since the analysis of one mental property requires invoking other mental properties and the reduction cannot be carried through. However, if such an analysis were available, it would be an example of the relation between two kinds of pleonastic properties. Nevertheless, other metaphysical relations that are discussed below are not just relations between predicates and therefore they cannot be taken as holding solely between pleonastic properties.

Before we come to the analysis of mind-body relations, a word or two about physicalism and dualism. My discussion is limited to common physicalist accounts. Borrowing a formulation from Braddon-Mitchell and Jackson (1996: 276), we may define physicalism as the view that the world contains nothing over and above the entities recognised by the physical sciences and their naturalist extensions.[18] Kim (1996: 10–12) is more specific, defining physicalism as an adherence to three principles: that the mental supervenes on the physical; that the mental is determined by the physical; and that there are no mental substances. I assume that a consistent physicalist should not allow that there are inflationary *mental* properties either. A physicalist would be committed to the claim that all inflationary entities are those that belong to the natural basis and that the natural basis consists of ontologically basic *physical* entities and their configurations. It is difficult to see how mental entities could fit into this class. As I argue in this chapter, there are ways of understanding several common mind-body accounts so that they do not presume inflationary mental properties. In case the account postulates irreducible and inherently mental and inflationary entities, the account should be rejected by the physicalist. In general, the urge to stick to inflationary mental properties comes from the unacknowledged and unjustified adherence to the principle Φ.

The ascription theory is physicalist in the sense of Braddon-Mitchell and Jackson, for it assumes that the natural basis is physical. As concerns Kim's constraints on physicalism, my account allows global supervenience, as I shall explain below, and rejects mental substances. And there is no issue of inflationary mental entities being determined by physical states of affairs, since there are no inflationary mental entities. However, it does not presume that the applicability of mental

18. There is no need to reserve 'physicalism' for type-type identity theory, as Block (1978) does. He therefore thinks that functionalism and physicalism cannot both be true.

descriptions can be explained in purely physical terms. Still, our basic ontology can be physicalist, for this explanatory failure is only a feature of our concepts.

I do not discuss substance dualism in this book, for I have nothing new to add to the common critique that is directed at it from various sides. The notion of mental substance has severe problems, if it is intelligible at all, and it is more important to centre on views that are more widely supported than substance dualism on the contemporary scene. With respect to property dualism, I do not object to the dualism (or, more aptly, pluralism) of pleonastic properties. There are as many kinds of pleonastic properties as there are kinds of predicates. However, I am opposed to the dualism of inflationary properties, and I touch on this in those places where I treat antireductionist accounts of the mind-body relation.

2.3 Supervenience

For many, supervenience seems to offer an opportunity to specify the content of non-reductive physicalism, which does not commit one to specific views about the nature of mental properties. In a nutshell, it can be used to express a dependence relation between mental and physical properties. It is said (e.g., Kim 1998: 9) that the class of mental properties μ supervenes on the physical property class π, just in case necessarily, for any property M in the class of mental properties μ, if any object x has M at time t, then there is a property P in π such that x has P at time t and anything that has property P at a time also has property M at that time. This formulation captures the notion of weak supervenience, which only holds within one possible world. Sometimes this is regarded as being too weak and puzzling. For it is consistent with weak supervenience that in other possible worlds, which are relevantly different to our world, P occurs without M occurring. It is only necessary that x would have *some* physical property, e.g., N, and anything that has N in that world also has M, but M might supervene on different physical properties in different worlds.[19] This makes the relationship between mental and physical properties far too puzzling (see Heil 1998b: 149). Therefore, a strong supervenience relation is postulated by adding the requirement that it is *necessary* that anything that has P also has M. This in turn is also problematic, for it undermines the point of antireductionism by making it possible to infer the possession of M-property from the possession of P-property.

---

19. This case differs from the situation in which M is multiply realised. In that case too, it does not follow from having M that one has a realiser P_1, for M could be also realised by P_2. However, it is still assumed that having any member from the set of realisers is sufficient for having M. That is, whenever one has P_1, one has M, and whenever one has P_2, one has M. In the situation discussed above, by contrast, in another world, one could even have P without having M.

There is an alternative formulation in terms of indiscernibility, which is generally accepted by those who accept the first formulation (Kim 1998: 10). It says that necessarily, any two subjects that are (add: 'in all possible worlds' for strong supervenience) indiscernible with respect to all *P* properties are also indiscernible in their *M* properties. This is sometimes also inverted in the slogan: "no mental difference without a physical difference".

It is often pointed out in discussions of supervenience that this notion is much too unspecific to ground antireductionism of the right sort. This is the view that although mental properties are something over and above physical properties, they are still dependent on the latter, while physical properties do not depend on mental properties. As stated, however, supervenience does not exclude the dependence of the physical on the mental. Neither does it exclude a mere covariance of mental and physical properties. It is also compatible with the scenario in which the possession of both properties depends on some third kind of property (cf. Kim 1998: 11). Although the supervenience claim says that physical properties are sufficient for mental properties, it does not contain the claim that mental properties are not sufficient for physical properties. The latter clause should be added to make the dependence asymmetric. Similarly, if one prefers the indiscernibility formulation, then it should be added that nothing about *P* properties follows from the indiscernibility in *M* properties. The case is rather that if mental properties of two subjects are discernible, then their physical properties must be discernible as well. It would be a proper objection to the supervenience thesis if there were a situation in which subjects are mentally discernible, but physically identical. But from the case in which subjects are mentally indiscernible and physically either alike or discernible, nothing follows about the supervenience thesis.

The crucial point, however, is that even when such amendments are made, and the formulation of asymmetric dependence is achieved, the relation still remains metaphysically unspecific. This is to say, supervenience is not a solution to the mind-body problem, for it itself stands in need of explanation (cf. Kim 1993c: 167; Heil 1998b: 150–151). There must be some other relation in virtue of which the asymmetric dependence between mental and physical properties holds. It might be identity or functional reduction or composition. In this sense, the supervenience claim is merely a reformulation of the problem rather than the answer. This point does not even depend on the pleonastic-natural distinction, for it can also be made without distinguishing these strata. The distinction only makes the case stronger. We can consider supervenience as pinpointing a relation between pleonastic properties. As such, it is consistent with there being just a *single* natural basis (for the given case) that is describable by both physical and mental predicates (hence pleonastic properties), and it is also the case that the application of mental predicates is asymmetrically dependent on the application of

physical predicates. The rationale for this dependence can be explicated either by reference to the nature of the natural basis or by the related application conditions of respective predicates. However, it is also compatible with a situation in which there are *distinct but correlated* natural mental and physical properties.

At this stage, it is instructive to introduce another notion of supervenience – ascriptive supervenience (see Klagge 1988). Whereas the previously discussed notions of supervenience are commonly taken to be ontological, holding between properties, ascriptive supervenience holds between kinds of judgements or ascriptions. It says that one kind of judgement about some object cannot differ without a difference in another kind of judgement. For example, Donald Davidson (1970, 1993)[20] has presented his supervenience claims as ascriptive rather than ontological.[21] Colin McGinn (1978) followed him in this usage, and endorsed the supervenience of mental judgements on behavioural judgements. Contemporary discussions usually centre on ontological supervenience and the ascriptive version is dismissed, perhaps as a relation between mere judgements. But as we can view ontological supervenience as a relation between pleonastic properties, then these two accounts do not seem that different after all. Note that ascriptive supervenience need not be committed to the claim that judgements about physical states imply judgements about mental states in virtue of their meaning. Neither should ontological supervenience be burdened with the claim that physical concepts somehow imply mental concepts in virtue of their meaning. However, both are consistent with the claim that the applicability of mental predicates to an object cannot change without a change in the applicability of physical predicates to this object. We may put this more exactly, using the Kimian notion of weak supervenience: the application of a mental predicate M supervenes on the application of a physical predicate P, just in case necessarily, for any object x and predicate M, if M is applicable to x, then there is a physical predicate P such that P is applicable to x and it is true of any object that if P is applicable to it, then M is applicable to it. Of course, there is a distinction between the applicability of a predicate and the judgement, but this difference is quite unimportant here.

Viewing thus supervenience as a relation between pleonastic properties or judgements, it becomes sharply perceptible that a story has to be told as to why should there be such a relation. In cases where the supervenience does not hold in virtue of meaning-relations between concepts, there has to be a metaphysical

20. To give a sense of the original year of publication, all references to Davidson's work are to his original essays and their page numbers.

21. However, Davidson's formulation of supervenience is importantly different from the standard Kimian formulation from which I proceed. When we come to token identity theories, I relate these to Davidson's notion.

ground for it. There is an extra complication here. A physicalist cannot simply claim that mental properties supervene on physical properties if both are understood in the inflationary manner.[22] If one assumes that in order to secure some sort of antireductionism, mental inflationary properties must be something over and above physical properties, then the resulting antireductionism is strong and mysterious. After all, what could natural or inflationary mental properties be? What makes them mental? And inherently and irreducibly mental as such? If they belong to the natural world, how can they still be something over and above physical properties? So, I think that it is much easier to make sense of supervenience-based antireductionism if it is presented as a claim about pleonastic properties. Of course, in that case it would be a much less interesting position than originally intended, for it would be compatible with a strong form of physicalism, which says that all natural properties (or all properties there are) are physical. It would be antireductionism only at the conceptual level, if coupled with the claim that mental terms are neither completely definable nor translatable into physical terms. Its claim would thus be that whenever an object has a mental description, it has some physical description (subject to certain constraints).

Thus far I have argued that supervenience claims need to be based on other mind-body relations and that mind-body supervenience, understood as holding between inflationary mental and physical properties, does not make sense for a physicalist. Supervenience, understood as a relation between pleonastic properties, remained standing. In the following, I reject the weak supervenience of the mental on the physical.[23] This rejection holds no matter whether supervenience is taken to be ontological or ascriptive. By showing that such supervenience does not even hold within a single possible world, I also reject strong supervenience, which is supposed to hold in all possible worlds. Given that supervenience is a weaker mind-body relation than type-identity, I have also set my face against the latter. As we will see, the counterexample to supervenience that I am going to

22. On Chalmers' (1996) view, the relation between mental facts and physical facts, taken inflationarily, is *natural* supervenience. In this case, the mental facts supervene on the physical facts in virtue of there being contingent psychophysical laws. In my view, even if we agree that the supervenience relation which is backed up by natural laws can hold between entities of the natural basis, the extension of such a supervenience relation to the psychophysical case remains deeply suspect. It would require a host of *ad hoc* psychophysical laws that would relate each mental fact or property to all those physical properties that realise them. If there is multiple realisation, such laws become excessively complex (cf. the criticism in Heil 2003a: 243–247). In addition, the antireductionism of the resulting view is extreme, for it is not even a version of physicalism anymore.

23. Note that the following thought experiment does not reject Davidson's even weaker notion of supervenience, which will be discussed in 2.4.

discuss is also troublesome for functionalism. I present further objections to these views in the next sections. At the end of this section, I discuss global supervenience and tentatively subscribe to a version of it.

There are two cases that would refute the supervenience thesis. They are both cases of mental difference with physical sameness. They can be distinguished by the order of entailment. In the first case, we have two physically identical (or relevantly indiscernible) subjects, but it does not follow that they are also indiscernible in mental respects, for they are mentally discernible. In the other case, we start from the mental difference, and failing to find a relevant physical difference, we end up with physical (relevant) sameness.

I think that common externalism provides the required case (Putnam 1975c; Burge 1979). Since the target is weak supervenience, we need a situation in which there is a mental difference in the form of a difference in thought contents even for same-world physical duplicates. The scenario that I am discussing to support externalism takes place within the same world; it does not presume travelling across possible worlds. Thus, imagine the following case: there is a person, Oscar, who believes that milk is a dairy product that is derived from cows, and that it is white and drinkable. Imagine that Oscar has a twin called Toscar, who also has similar beliefs about milk. Now imagine that there is a country on Earth that is in much better shape in industry and chemistry than in agriculture. It is called Milkonia and it is located somewhere in Eastern Europe, having inherited a strong industry and weak international relations from the former socialist time. There are no cows in Milkonia. Since it is not a member state of the EU yet, there are no export-import relations concerning soft drinks and dairy products. There is therefore no milk in Milkonia. However, there is a product, ®-milk, that tastes and looks like milk, but is actually produced by the local chemical factory and it has absolutely nothing to do with cows. Moreover, its chemical structure is as different from the structure of milk as H_2O is different from XYZ. Think of it as UHT-milk in the extreme – to the point of its not being milk anymore. Now Toscar pays a visit to Milkonia, and drinks a glass of ®-milk in the local bistro. He knows nothing about the peculiarities of the Milkonian ®-milk industry and therefore he can be ascribed the belief that he drinks milk and the thought that milk is white. His twin, the physical duplicate Oscar, sits in a local pub and drinks real milk. The belief that he drinks milk and the thought that milk is white is ascribable to him. Now, an externalist would say that Oscar's belief is true, while Toscar's belief is false, for Toscar drinks ®-milk. An externalist would also say that the content of Toscar's thought is about ®-milk, and the content of Oscar's thought is about milk. In this respect, Toscar's thought is not the same as Oscar's thought. Hence, there is a difference in the twins' thought contents, but as presumed, they

are physically relevantly similar. Hence, there is a difference in mental properties without a difference in physical properties. Similarly, there is a difference in the applicability of mental predicates without a difference in the applicability of relevant physical predicates. In this scenario, one does not infer from physical sameness that the twins should be mentally alike. And if one accepts that their thoughts differ since their contents differ, then it does not yet follow that there must be a relevant physical difference between Oscar and Toscar.

It can be objected to this story that even if the content of mental states does not supervene on physical properties, there is still some "narrow content" (Fodor 1987a) or some "local causal surrogate of content" (Crane and Mellor 1990: 194) which is identical among twins. I do not object to this answer, but if that much is conceded, then it is also conceded that the supervenience thesis does not hold for all mental properties, namely for "wide" intentional properties.[24] Then the physicalist owes an explanation as to how content-properties are related to physical properties. This is an additional task that remains even when the intentionality problem, or the mind-world issue as specified in Chapter 1, is solved. It is the more general worry about the place of abstract objects in the physical world. Usually it is thought that it is narrow content that plays the causal role and therefore the external, non-supervening differences do not really matter.[25] I postpone the discussion of the severe problems facing the supervenience thesis in the field of causation until Chapter 7. Here I would like to note only that those external differences might well have a causally explanatory role and thus can still be causally relevant.

The foregoing was an objection to supervenience from the standpoint of common externalism. This is a controversial position, and its implications are still widely disputed (e.g., Segal 2000). However, given some qualifications, I share the general externalist view that external factors matter in the individuation of

24. This scenario is a problem for functionalism no matter how we construe functional properties. If we individuate functional properties narrowly as proposed by Fodor, then the twins have same functional properties, but different content properties. Hence, content properties do not even supervene on functional properties. If we individuate functional roles widely, twins come to have different functional properties, but there are no differences in their brain states. Hence, functional properties do not supervene on neural properties. Although functionalism says that functional states are multiply realisable, it is still true that if one has a realiser, one has a functional state. In the present example, the brain state belongs to a set of realisers, but apparently it does not suffice to realise the same functional state in Toscar and Oscar.

25. For instance, Jerry Fodor (1987a) has argued that only narrow content matters for psychology, since it is individuated by causal powers that supervene on the causal powers of brain states.

thought contents.[26] According to the ascription theory, the possession of mental properties (still in the pleonastic sense) does not carry any strong implications about the agent's internal physical properties. Which mental properties one has depends solely on which mental predicates are ascribable to one. This is in turn determined by various factors, among which are ascription sources and the ways in which these sources hang together. There is also variance in determining mental properties, given that there can be more than one equally acceptable way of combining information from various sources that yield mental ascriptions. Hence, it does not make sense in this model to single out a physical property which is sufficient for the possession of a given mental property. Even physical duplicates could have different mental properties given that they are located in different environments and different schemes of ascription are applied to them.

At this point, however, it seems natural simply to expand the supervenience base. For example, common externalism rejects the narrow supervenience of mental content on the properties of brain states, but it would be consistent with externalism to say that content supervenes on both brain states and the environment. In that case, for every mental difference, there is still some physical difference, if not in the brain, then in the context.[27] However, when we make this move, then we may already be taking a step towards interpretivism, for interpretation may be involved in fixing the context that determines content. Otherwise, externalists need to maintain that context determines content by itself, brutely, via some uninterpreted causal or teleological relations. This is the line many have taken, and I am not going to argue with it here. It is perhaps needless to say that such accounts are controversial and we still have no clear sense of how a mere context (abstracted from other people, interpreters, language-users) could have a causal or teleological effect on what we mean or believe. The question I am turning to instead is whether the ascription theory is consistent with the supervenience of the mental on the physical, provided that the supervenience base is enlarged.

There is no straightforward rule for deciding what should be involved in the context that is to be included in the supervenience base. However, there is yet another notion of supervenience that could prove useful for fuelling our intuitions. It is the notion of *global supervenience* (e.g., Kim 1984a: 168; Jackson 1998: 11–12)

---

26. In my view, the content of beliefs in general should be specified in those terms that the subject can acknowledge or would use in the self-ascription of this content. Hence, although the environmental factors guide the specification of content, other considerations may defeat their proposal. In any case, the ascription theory leaves open the possibility of ascribing different contents by the same words like 'drinking milk'. The issue of how to individuate propositions is separable from my general task of giving an interpretivist account of thought, and won't be pursued in this book.

27. E.g., William Child (1994: 76) argues that one can interpret molecularly indistinguishable subjects differently only in virtue of differences in context.

that is put in terms of whole worlds. It says that the set of *M* properties globally supervenes on *P* properties just in case any two *worlds* that are duplicates with respect to *P* properties (that is, are indiscernible in respect of their *P* properties) are also duplicates with respect to *M* properties. Applied to physicalism, it results in the claim that if there is a world indiscernible with respect to all of its physical properties from our world, then these worlds are also indiscernible with respect to all of their properties.[28]

The appraisal of global supervenience has to take into account whether it holds between natural properties or between pleonastic and natural properties. If global supervenience is taken only as a claim about natural properties, then I agree that if there is another world, which has all natural properties in common with our world, then they are duplicates with respect to all natural properties. This sounds trivial, given my naturalist assumption that in any serious ontological sense, there exist only natural properties. Perhaps global supervenience becomes interesting only if one assumes that there is more than one kind of inflationary property, some kinds thus being more basic than others. Moreover, some of these kinds have to be non-natural, being merely supervenient on natural properties. It does not matter if there are just different kinds of natural properties, for they would be included in the supervenience base that contains all natural properties. The issue is trickier when we take global supervenience as the claim that if there is a world indistinguishable with respect to its natural properties from our world, they are indistinguishable with respect to their pleonastic properties as well. On the one hand, there is a temptation to say that there is no guarantee that the same natural property should be described in the same way in different cases (and hence correspond to the same pleonastic property). On the other hand, however, now we are considering the whole world, and one might be tempted to conclude that if we pair all natural properties with all possible descriptions that could be given to them, we shall get a dependence between these two large sets.

Let us consider this from the point of view of the ascription theory and limit our discussion only to the global supervenience of the mental on the physical. The central claim of the ascription theory is that having a mental property is constitutively dependent on whether this property is ascribable to one. Hence, in a sense, mental properties supervene on ascribability. If we look more closely at what is

28. Jackson (1998: 12–13) argues that physicalism should consist in an even weaker claim, namely that "any world, which is a minimal physical duplicate of our world is a duplicate simpliciter of our world", in which a minimal duplicate world is a world that is physically like our world, and contains no additional entities. If the last clause is omitted, then a physicalist is required to disallow the possibility of a world which is a physical duplicate of ours, but contains additional non-physical entities, but this might be too strong. However, for our purposes, this distinction can be overlooked.

involved in the ascription, we see that the ascription sources are not limited to pleonastic physical properties and facts. They include also social facts, as well as other intentional and mental pleonastic facts. It is assumed that all these pleonastic facts and their relevance appraisals, accompanied by some way of combining them to get a constitutive ascription, will tell us which mental properties are ascribable to the subject, and given the interpretivist assumption, would hence yield the facts about the subject's mental states. The physical and intentional facts themselves thus underdetermine what is ascribable. One has to add a measure of relevance to these facts, as well as to decide how to put this all together to maximise the harmony of ascribed states and contents with the evidence for ascription.[29] There is no objective instrument for the appraisal of relevance of the ascription sources; its specification has to be bound up with a potential ascriber. The notion of a potential ascriber is ineliminable from this account. For it is only through the latter that some facts can be taken as intentional and social.

It is thus indeed the case that if we duplicate a world, which contains all pleonastic physical, social and intentional facts or properties with their relevance assignments plus all those ways of combining these facts, then the duplicate world does not differ with respect to pleonastic mental properties. However, this sort of supervenience base is highly non-reductive. It contains intentional and social facts and the mention of a potential ascriber. The attempt to specify these facts in physical terms does not succeed in breaking out of the intentional circle, since it would presume further intentional facts. There are no one-to-one bridge laws either to connect intentional facts with the purely physical supervenience basis. Although these intentional facts are pleonastic, there is no way to bring them down to non-intentional descriptions, at least not in the ascriptionist approach.

What can be concluded about the global supervenience of pleonastic mental properties on natural properties? There are, as it appears, two ways to read this sort of global supervenience. On the first reading, we take our supervenience base to consist in *pleonastic physical properties* (that is, predicates which play a role in physical theories, which one might take as privileged descriptions of the natural basis). It follows from what was said in the previous two paragraphs that mental properties do not supervene on such pleonastic properties, for their supervenience basis contains also pleonastic properties that are not physical. Fixing only pleonastic physical properties does not yet give us pleonastic mental properties. However, on the second reading, we fix the *natural basis or properties* (no matter how

29. It should be noted that the number of mental properties ascribable to a given subject in a given situation might be quite large. Even if all the facts are in, one need not get a single mental profile for the subject. One might get several different mental profiles, which are all equally well ascribable (given the standard outlined in Chapter 5), since the same ascription source can get different relevance assignments, thus yielding different possible results on what is ascribable.

they are described) and see if the pleonastic properties are fixed as well. Indeed, they should be, if we take all natural existents to exhaust all that is the case as well as to determine all that could be the case. On my account, all mental properties are only pleonastic and ascribing minds does not add new natural objects to the world: it is just fiddling with descriptions. However, even if one agrees with the global dependence of mental descriptions on the natural basis, it is still an open question in what sense and manner the natural basis determines all these pleonastic properties and facts. It is especially so in view of the fact that these are just descriptions (and in most cases potential descriptions), hence abstract. Although they are manifestable, they need not always be actually tied down with particular manifestations (marks of ink, movements of vocal cords, etc.). There is still no generally accepted account of how abstract descriptions, quantities, etc. supervene on the natural basis, which itself need not have a privileged description.[30] In this regard, subscribing to the second reading of the global supervenience claim is nothing more than an article of faith in naturalism.

The upshot of the present section is that even the weak supervenience of mental properties on the neural properties of the subject is not tenable. Still, one may suggest, perhaps mental properties supervene on physical properties if context is included in the supervenience base. However, this is not quite compatible with the ascription theory, the externalism of which is more resolute. The ascription theory is compatible with global supervenience when the supervenience base is taken to contain the whole natural world, but not when it contains only pleonastic physical properties. In the pleonastic case, mental properties supervene on ascribability, which presumes pleonastic intentional and social facts.

2.4 Reduction and identity

Let us move from discussing weaker or non-reductive mind-body relations to considering reductive physicalism. The bulk of this section is devoted to identity theories. Since mind-body identity is closely connected with reduction, I shall begin by discussing the latter notion. It is customary that accounts of reduction deal with the relations between whole disciplines. For example, it is asked

30. The second case thus differs from the currently debated issue of whether physical descriptions or truths *entail* (either a priori or a posteriori) all other descriptions or truths (e.g., Jackson 1998, 2003). If the notion of entailment does not differ from supervenience, then the entailment question corresponds to my first reading of the global supervenience claim. If it used in some metaphysically stronger sense, where one set of facts or truths entail others, the notion is unclear and the accompanying metaphysical picture is hard to accept. For it seems to presume both that it is a determinate matter which set of truths is 'physical' and that the set is fixed independently of the contingent developments of our physics.

whether psychology is reducible to neuroscience. Putting the issue in these terms apparently originates from positivist and post-positivist discussions concerning the unity of science. This legacy is also evident in the fact that, at least for philosophers of mind, the familiar model of reduction is still the Nagelian model of theory reduction and the issue is often put in terms of laws and vocabularies. It is this notion of reduction that most antireductionists oppose themselves, if the notion of antireductionism is specified in the context of theoretical reduction. Given that there are other models of reduction, we get a peculiar situation in respect of functionalism, which is antireductionist in the Nagelian sense and perhaps in some further loose sense, but which itself is still a kind of reduction, namely functional reduction (cf. Kim 1998). I shall return to this topic in the next section.

The standard interpretation of the Nagelian reduction (Nagel 1961; cf. Kim 1996) is the following. The problem concerns the reduction of one theory T_1 to another theory T_2. In Nagel's model, this is accomplished if the laws of the reduced theory, T_1, can be logically deduced from T_2. Given that T_1 and T_2 have different vocabularies, there is a need for additional "bridge principles" that would provide for every primitive term of T_1 a corresponding term in the vocabulary of T_2. Assuming that M is a psychological predicate of T_1 and P is a neurological predicate in T_2, the bridge principle or law takes the following biconditional form: $(\forall x_1)...(\forall x_n)\,(M(x_1,..., x_n) \leftrightarrow P(x_1,..., x_n))$. Those bridge laws were supposed to represent contingent and empirical correlations between predicates that are nomologically coextensive. It was not assumed that M predicates and P predicates are analytically connected in virtue of their meanings. However, given the bridge laws, terms and claims of T_1 can be translated into the vocabulary of T_2. In a nutshell, a bridge law gives us a correlation of two kinds of predicates, which as a matter of law have the same extension. That is, it has to be a law that those sets of individuals to which these predicates apply are equal.[31] In other terms, we may say that a bridge law relates two kinds of pleonastic properties which are possessed by the same set of natural entities.[32]

31. It is important that the correlation is *nomological*, for otherwise one can generate counterexamples by adding to the right-hand side of the biconditional some terms with an extension containing nonexistent entities. But since ($M \leftrightarrow P \vee$ being a unicorn) is not a law, the coextensionality of P and ($P \vee$ being a unicorn) is no counterexample to this account (cf. Putnam 1969: 237).

32. Note that my notion of the natural basis differs from the notion of a (first-order) predicate's extension, which contains objects or individuals to which the given predicate applies. The natural basis of a (first-order) pleonastic property is not restricted to objects; it can also contain properties and powers, which are not objects in the relevant sense of the term. In addition, depending on one's ontology, what belongs to the extension of any arbitrary predicate need not yet amount to what one is prepared to take as the natural basis.

The Nagelian account has several problems as an account of reduction. It is unlikely that it characterises the process of reduction in science correctly. To mention just one point, familiar from the philosophy of science: often the old theory needs correction or even replacement, and hence Nagelian bridge laws are even undesirable, since they would attempt to connect an old erroneous theory with the new theory. There have been various attempts to rectify or replace the model of reduction. I will not go deeply into this, for developing a satisfactory account of reduction in science is not central to the present project. What I want to discuss instead is the relation of reductionism in Nagel's sense to other physicalist mind-body relations.

In general, there are three kinds of reduction: conceptual reduction, theoretical reduction and ontological reduction. Conceptual reduction is an analytic project of expressing the statements of one kind in terms of statements of another kind. Its main task is to give an analysis of concepts in terms that do not involve the concept. Usually it is said that such an analysis explicates the meaning of a concept by displaying its necessary and sufficient constituents. Most of the mind-body relations, with the exceptions of logical and analytic behaviourism and analytic functionalism, are not reductionist programs in this sense. The reduction of theories with the help of bridge laws is an example of the second kind, theoretical reduction. The third kind, ontological reduction, aims at ontological economy: it attempts to reduce the number or types of entities. In one sense, this is the metaphysician's project of reducing one entity-kind to another, for example, universals to particulars. This does not concern us in the present section. In another sense, this exhibits the urge to show that the world consists of entities of only one, e.g., physical type (this leaves open the possibility that there may be physical properties, powers and particulars). To put it in other way, not necessarily the way I would endorse, ontological reduction attempts to show that the truth of certain statements does not require the existence of facts of a certain type, and that they can replaced by another type of facts. For example, the truth of statements about mental states does not require the existence of mental facts or events; they are replaceable by certain physical facts or events. The main mind-body theories are examples of ontological reduction. The question is now whether we can use the Nagelian account of theoretical reduction to illuminate the ontological question. Many philosophers have apparently thought so, but there are problems.

Some of these problems, relevant to the present discussion, are repeatedly raised by Jaegwon Kim in various writings (e.g., Kim 1998, 2003b, 2005). The main problem is that even if the Nagelian account is acceptable as a model of how one theory reduces to another, it fails to reduce properties in the ontological sense. Assume that bridge laws help to deduce all statements of the target theory. However, what is not explained is the bridge laws themselves (Kim 1998: 95–96).

They were introduced in order to provide links between the target of reduction, T_1, and the reducing theory, T_2, and they might succeed in providing the 'translation manual' for the reduction. But if one's aim is ontological reduction, this is just one half of the task. The other half would be explaining why bridge laws hold. One might take bridge laws as brute facts, but this is not satisfactory, for a bridge law is nothing more than a contingent correlation and thus there should be a more basic story about why there is such a correlation. This case is similar to the one discussed in respect of supervenience. As in that case, bridge laws are compatible with various mind-body relations as long as they imply lawful correlations. This point is stressed by Kim (1998: 97), who notes that Nagelian reduction of psychology to physics is even consistent with substance dualism. Kim's worry is that the correlation set forth by bridge laws is too weak, since the predicates in a bridge law stand only in a contingent biconditional connection, which makes it possible that the corresponding properties are distinct. In terms of my scheme, not only are there two pleonastic property types, but there can even be two inflationary property types whose instances stand in perfect but contingent correlation.

One might come to think that theory-reduction is all there is to reduction if one assumes that there are no theory-independent facts or properties. In this view, all properties are incorporated into theories via the corresponding predicates and hence all issues concerning the reduction of properties are to be solved by the reduction of theoretical claims containing predicates. However, even in this picture, the same problem remains. The Nagelian model is inadequate, for there remains the task of explaining what reason we have for assuming that predicates of different theories apply to the same individuals. For given parallelistic dualism, it can still be the case that all persons, describable by a given mental predicate, are (by nomic necessity) also describable by a given physical predicate. But in the case of parallelism, this is so in virtue of the perfect correlation between two substances. This is an option that any reductionist should find disturbing.

One suggestion that is made to strengthen the case for reduction is to think of the relation between mental and physical states rather in terms of strict identity than bridge laws (Sklar 1967, more recently by Block and Stalnaker 1999). In this way one would retain the broader Nagelian model of reduction; only the correlation between predicates would be replaced with identity. The mind-brain identity theory was originally propounded by Place (1956), Feigl (1958) and Smart (1959). They explicitly distinguished their identity claim from the mere conceptual claim. If one says that a mental state, e.g., being in pain is identical to the state of c-fibres firing, then this is not a claim about the sameness of meaning of these words or the synonymy of these concepts. In terms of the general metaphysical scheme of this book, it is not the case that a pleonastic mental property

(or state) is identical to a pleonastic brain property (or state).[33] The identity involves a relation between two strata: it says that both pleonastic properties have the same natural basis. Therefore, it is potentially misleading to say simply that properties (or states) are identical.

It was the contention of these identity theorists that such identities are discovered empirically, and since at that time the general assumption was that there is no distinction between a posteriori and contingent, they claimed that the mind-brain identities are contingent. After Kripke (1980) separated these notions and made a posteriori necessities intelligible, it is generally conceded that statements of identity are necessary, given that the terms are rigid designators. Consequently, if Block and Stalnaker (1999) suggested replacing bridge laws with identities, they had necessary identity in mind. This makes the relation between mental states and brain states stronger than correlation and helps to make sense of the reduction as an ontological enterprise. It allows us to say that there are fewer entities in the world than it seems, namely, that there are no inflationary mental entities. There are only brain states in their place. Compatibility with dualism is thus out of the question and, since identities cannot be explained, one cannot ask why pain is c-fibre firing (cf. Kim 2003b: 570). In this regard, reduction by identities fares better than reduction by bridge laws, but this is only due to the fact that an ontological premise – namely, identity – was added to the account of reduction.

The notion of the mind-brain identity needs further clarifications. Identity as it was outlined above is an identity between types. Usually, this is put in terms of a given mental property-type being identical to a given physical property-type. This is contrasted with the token-identity thesis, in which case every token mental property m from a certain type M is identified with some token physical property p, but the identity of types does not follow, since different mental tokens of type M can be identical to tokens from different physical types.

It is instructive to put these claims in terms of pleonastic properties and the natural basis as well. For this, we need the notion of a pleonastic type in the sense that M (e.g., being in pain) is a pleonastic type in contrast to the pleonastic token of having m_i from the type M at time t. There is a natural distinction between the concrete instance of a property and the property as such (or a set of all its instances). The former can be viewed as a token property and the latter is a property-type. Of course, if properties are universals and in each property-instance there is a universal wholly present, then there is no difference between type and

33. The claim of identity can be put both in terms of properties and in terms of states. Identity of states requires identity of properties, for we may take *x*'s being in a state as *x*'s having a property (Heil 1998a: 76). For an alternative view, namely that states differ crucially from properties, see Steward (1997).

token identity.[34] Similarly, on the property-exemplification account of events, which is proposed by Kim (1976), events are identical only in case they are the instantiation or exemplification of the same property by the same object at the same time. Since this requires property identity, token identity entails type identity and, as Kim (1996: 60) himself puts it, "there is no interesting distinction between token physicalism and type physicalism on the property exemplification approach to the nature of events." So on those accounts, these two identity theories need not be distinguished. But on other accounts of property instances, there is a difference. It is a general contention that token identity is weaker than type identity, as it does not entail the latter. Usually, the identity of tokens is put in terms of events that are taken as particulars, unlike Kimian events, for which several people would rather use the term 'fact' (see Mellor 1995: 133; Crane 2001: 159). Expressing the token identity theory in terms of events presumes that properties themselves are types, and that property-instances are events. However, given my deflationary notion of property, there cannot be any harm in talking about token properties and type properties.[35] I see no obvious advantages in confining the token-identity claim only to events by definition, especially given that there is an alternative view which says that each instance of a mental property is an instance of a physical property, while the types of which they are instances are not identical. But I submit that this difference does not loom large.

Having this in the background, we can formulate the statements of type and token-property identity theories. They attempt to express the ontological assumptions or commitments of these theories and it is not assumed that all theorists themselves have stressed these commitments, or, at least, clearly not in these terms. Obviously these are not meant as alternative accounts of the logic of identity relations, nor should they be understood as identity theories about pleonastic properties themselves. That would be a use-mention confusion. Instead, I utilise the notion of a pleonastic property to express what is assumed, when it is said, e.g., that mental properties are identical to physical properties. The same applies

34. Applied to my formulations below, this is to say that the natural basis consists of a universal, not that two universals correspond to the same basis-universal.

35. If properties can also be viewed as tokens, and not only as types, then this is a step towards severing the connection between being a property and being a type, at least at the inflationary stratum, which is a good thing to my mind. My talk about property-types leaves it open whether the type is itself a universal or the set of its instances or the set of similar instances at the inflationary stratum. It would even be an open question whether the type-property itself is an inflationary property, for it is possible to hold that there are inflationary tokens of properties (tropes), while their belonging to the same type is only a matter of applying a certain description to some tropes (Heil and Robb 2003). On such a view, all types would be merely pleonastic.

to my later formulations of functionalism as well. (These formulations are general, but they apply to the mind-body relation, if '*M*' or '*m*' are read as mental, and '*P*' or '*p*' are read as physical, or neural, to be specific.)

Type-type identity theory: For each pleonastic *M* property-type, there is a pleonastic *P* property-type, such that both have the same type of natural basis (the same type under 'P' description).

Token-token identity theory: For each pleonastic property token *m*, there is a pleonastic property token *p*, such that both have the same token of the natural basis (the same token under 'p' description).

Talking about a pleonastic property having a natural basis does not bring any mysterious having-relation to bear. As I have already pointed out, natural properties and pleonastic properties are not properties in the same sense, and one having another as a basis is not an obscure metaphysical relation. It is just a more compact way of saying that we apply certain predicates to natural entities, and given the deflationary scheme, we get pleonastic properties from these predicates by the something-from-nothing transformation. It should also be obvious from the foregoing that when I say that a pleonastic property has a natural basis, it need not mean that the pleonastic property is realised by this basis. I will discuss the realisation relation in the next section. Finally, one may also wonder whether it makes sense to talk about one token having another token as a basis. But given that these tokens are pleonastic, all this means is that for each true case of predication, there is some concrete instance of a property, which makes the predication true.

There is no need to assume that the corresponding predicates must belong to any particular language or that they must be known to us. Identity theory is not an epistemic thesis. However, what we need is the prospect of finding an appropriate physical description, no matter how complicated, for otherwise it would not make sense to say that the natural basis is physical. That the sameness of the natural basis is specified in physical terms in these formulations is meant to express the primacy of the physical in determining what counts as natural. It puts an accent on the sense that, for some, only the physical description succeeds in "limning the true and ultimate structure of reality".[36]

---

36. This famous quote is from Quine's *Word and Object*: "If we are limning the true and ultimate structure of reality, the canonical scheme for us is the austere scheme that knows no quotation but direct quotation and no propositional attitudes but only the physical constitutions and behaviour of organisms." (Quine 1960: 221.)

These formulations include the two-strata aspect that is implied by the main mind-body relations. If one comes to count how many things (in this case, natural properties) there are in the world in a given problem-domain, then in case of type-identity, there is only one type, and in case of token-identity, there is only one particular token in each instance. My formulation aims to make this clear. The identity theory is not a reductive claim at the level of language usage or meaning, hence in both cases there remain two different predicates (or pleonastic properties). Identity is found at the stratum of the natural basis. What makes the relation one of identity rather than a correlation is the fact that each type (or a token) of the natural basis is necessarily identical with itself. Down at the stratum of natural properties, if the identity is at issue, there is no room for correlation, no matter how tight that may be. As we saw earlier, the reduction by a bridge law biconditional is, by contrast, compatible with the nomological correlation of two distinct natural types.

It may be protested that my formulations do not convey the gist of identity theories correctly. For an identity theorist might want to say that there really are mental properties in the inflationary sense, and this is so just due to the fact that mental properties are identical with physical properties. However, if we reflect on this, then it becomes evident that if a property is identical with another property, then they really are one and the same property, and we have one property, not two. If an identity theorist insists on taking mental properties as inflationary, then one is bound to say that there are two separate inflationary properties, which are nevertheless identical. Note that this is not the view that there are separate instances of the same property. Rather, it should be understood in the following way: as applied to types, it would insist on saying that there are separate and identical types; as applied to tokens, it would insist that there are separate inflationary tokens, which are nevertheless one and the same token. It is hard to make sense of such a view.

In analysing the identity theories, I shall begin with a discussion of token-token identity theory. Quite often, token physicalism is put in terms of the idea that there are only physical particulars, hence all particulars are identical with physical particulars (e.g., Stoljar 2001). Apparently, expressing token identity theory in terms of the identity of particulars is due to the treatment of Davidson's account as exemplary (see Davidson 1980a). Indeed, in his view, all events (which are particulars for Davidson) are physical, but some have mental descriptions as well. The difference between this account and my formulation above lies in the fact that I intended my definition to apply also to accounts that are framed in terms of properties. The Davidsonian view is centred on those particulars that have properties, but there are also accounts that focus on property

instantiations, tropes[37] or property exemplifications (cf. Kim 1998: 121). In my terms, the identity of particulars would be the following view:

Token identity theory for particulars: For each pleonastic particular *m*, there is a pleonastic particular *p*, such that both have the same token of the natural basis (the same token under 'p' description).

We get pleonastic particulars whenever we describe something as a particular. The formulation above leaves it open what the more specific nature of the natural basis is, e.g., whether its particularity can be further analysed or not. If we read my former token-token identity formulation by assuming that the instance of a property just is a particular event or that the instance of a property is a particular, these formulations can be reconciled. The difference arises whenever we assume some other account of token properties. The common core of both views is that neither of them implies type-type identity theory (given the aforementioned restrictions concerning some eccentric views on property instances). Each token of a mental property has a physical basis, but nothing is said about mental types. Similarly, in Davidson's account, the identity of events does not imply anything about the relation of types.

How might one assess the token identity view? For brevity, I phrase my claims mainly in terms of the identity of particulars or events, but the same points apply, *mutatis mutandis*, to the property instance-view as well. It says only that in each instance of having a mental token, there is only one relevant token property or particular at the natural stratum. It is a weaker position than type-type identity, for it implies nothing about types. Token identity is also a weaker relation than Kimian weak supervenience, which states that within a world, a mental property *M* supervenes on a physical property *P* such that whenever one has an instance of *M*, one has an instance of *P* and it is true of any subject in that world that if this subject has an instance of *P*, it has an instance of *M*. But all that token identity commits us to is the view that whenever a mental event occurs, there is *some* physical event to which this mental event is identical. It does not ground

37. Tropes are also particulars, abstract particulars, but here the crucial difference is that the Davidsonian identity theory centres on objects that have properties, and not on the identity of properties, while tropes are in general viewed as particularised properties. However, in a sense, my two statements of token identity theory are equivalent if one takes token properties as tropes and tropes as particulars. But see also Robb (1997: 187) who distinguishes trope monism from Davidson's monism. According to trope monism, mental and physical types are distinct, and mental events are physical events, but in addition, every mental trope is identical with a physical trope. In Robb's view, this helps to avoid the epiphenomenalism of the mental. For a dissenting voice, see Noordhof (1998).

the second conjunct of the supervenience thesis, namely the claim that for every subject, whenever there occurs a token physical event p from the same physical type P, the subject has a token mental event m from the same mental type M. The reason for this is that it can be the case that in other subjects, a particular m from the same type M is identical with a physical event from a different type. Or in case the subject is the same, it can be that at different times, another token m from the same mental type M is identical with a token from the different physical type P_1. It follows that even if weak supervenience fails, token identity theory can still be true. In light of the previously discussed externalist objection, an advocate of token identity theory can consistently maintain that although there is a mental difference between identical twins, Oscar and Toscar, it remains the case that when Toscar thinks that he is drinking milk, there is still some physical event going on in his brain which can be identified with his having this particular belief. It is intriguing to note that Davidson's own formulation of supervenience is as weak and unspecific as his token identity claim. Davidson (1985b: 242) writes: "a predicate p is supervenient on a set of predicates S if for every pair of objects such that p is true of one and not of the other, there is a predicate of S that is true of one and not of the other." Unlike Kim's notion of supervenience, there are no further constraints here on the predicates of the S set.

But since the token identity and Davidsonian supervenience only make such weak claims, they are not especially interesting. As explained above, I assume that at the natural stratum, there are neither mental particulars nor mental property instances. If one has mental properties only in the pleonastic sense, then having a mental property is not a matter of instantiating a mental universal: it is just a particular case of falling under the mental description. The same applies to mental particulars. I find any other notion of mental particulars unintelligible for a physicalist if not downright incoherent. Indeed, what could it be that is both particular and inherently mental? Those who do not agree with this way of understanding mental particulars must give an alternative account of the inflationary mental token which is *essentially mental* and, at the same, identical with some physical token. I cannot imagine any such account, other than the descriptive approach I sketched, but this leaves no room for inflationary mental tokens. In this sense, I agree with Davidson's claim that an event is mental whenever it has a mental description. Davidson himself is an interpretivist about the application of mental predicates and holds onto monism at the level of token events. In the light of what I shall say below, however, I see no reason to follow him in his identity theory.

I am suspicious about the claim of identity. It is undeniable that in each case of successful prediction or proper ascription, there is something going on in subject's brain or equivalent physical system, whatever that might be. It may seem to some that interpretivism does not imply anything about the physical structure.

One might say that it is open for us to ascribe mental states to non-physical subjects such as souls or God, etc. However, to say the least, such ascriptions would be spurious, for they would lack any ground whatsoever. There cannot be any physical evidence to guide the ascription of thought to non-physical creatures. This can only be done in fiction. Ascribing beliefs to fictional characters does not count here, since in those cases the ascription sources would also be part of the fiction. I assume thus that the real subjects of ascription need to have some physical constitution, for otherwise it would be mysterious how they can sustain complex behavioural patterns that merit the ascription of an intricate network of mental states. However, I see no reason to *identify* mental events with physical events (cf. Dennett 1994: 534–535). Of course, if one inflates mental properties, then one feels compelled to domesticate or legitimise them in some way, for otherwise one would be faced with some ontological loose ends. In other words, if one holds that there must be something that makes all true mentalistic ascriptions true and assumes that the world consists of events or facts (in the inflationary sense), then it is easy to conclude that such ascriptions must be made true by mental events or facts. And if one happens to be a physicalist, then the conclusion must be that, strictly speaking, mentalistic ascriptions are made true by physical events. Hence, if there are mental events, they must be identical to physical events. However, if one takes the deflationary approach, there is no need to legitimise mental descriptions in this way. The story that should be told about the applicability of mental predicates need not invest in metaphysical relations.

Viewed from this perspective, there is no pressure to legitimise mental entities by identifying them with some natural entities. It is not even obvious what to make of the claim in the deflationary context. For identity, it is not enough to say that whenever some mental predicates are applicable to *x*, some physical predicates are applicable to *x* as well. Mere ascriptive supervenience does not give us the token identity yet. The claim must be that the application of a mental description is true of some token brain event. However, this goes against the common usage of mental terms. It makes more sense to say that mental predicates are applied to the whole person and not to some physiological parts of its body. It is true that neuroscientists might occasionally ascribe content and function to certain parts and processes in the brain in order to link their activities to abilities specified in folk-psychological terms. However, it does not follow from this that each common-sense application of mental predicates is a claim about some event in the brain. In other words, mental terms are mostly applied at the personal level and not at the sub-personal level. According to interpretivism in general, the application of mental predicates is highly dependent on historical and environmental factors. One would thus need some substantial arguments to support the view that mental events are identical to some local and narrow

brain events. If they were to be identified with anything at all, then it should perhaps be with some complex events comprising historical, environmental and neural features. But even in that case it makes little sense to say that each application instance of mental predicates applies to a particular event comprising both some neural states and environmental slices. Such a token would be spread indefinitely around in social and physical space as well as lacking a definite location in time, since there is hardly any good way to demarcate the relevant portion of a subject's past history that is relevant to the specification of his mental states. Hence, the putative natural event that is supposed to make true an ascription of a mental predicate would be an event without a specific location in either time or space. If events are individuated by their spatio-temporal locations, then to say that such complex and extended entities are events would stretch the notion of an event. (The same point applies also to the instances of such neuro-historico-environmental properties, which require the exact place and time of their instantiation as well). If events are individuated by their causes and effects, it is hard to see how can such a nebulous entity bears causal powers. (See Baker 2001: 213 for similar considerations.)

Note that the token identity theory does not give us any clue about the events in the brain with which mental events are identified. It is not presumed that we must know which brain events or physical token properties these are. Since the particular brain event can also be taken as comprising the state of the whole brain at a time, the theory has no immediate neuropsychological consequences. Since the view is so unspecific, its explanatory value is minimal. The token identity theory makes only the ontological point by formulating a version of monism and rejecting the identity of types. Indeed, if the identity of types is too strong, then token identity is a good monist option. But as stated, given the physicalist or naturalist commitment that there are no natural and purely mental entities anyway, there does not seem to be any additional need to identify tokens. Using the notion of identity helps to express the idea that all there is to a mental event or to the instance of a mental property is some physical particular or token property. On the deflationary view, this still makes little sense, for the application of a mental description is nothing of the sort that needs to be exhausted by or based on a physical particular, nor is it intelligible how it could be.

It might be objected that the token identity view has the advantage of realism concerning the mental. For given that every mental event is identical with some physical event, irrealism concerning mental events is blocked. One may also think that the identity view is useful in explaining mental causation, for given that mental events are identical with certain spatio-temporal particulars, they have all the causal powers these particulars have. I am not moved by this objection, since this works only if there are such definite spatio-temporal particulars that can be

found among internal events in the brain. But this runs contrary to externalist intuitions. Moreover, in my view, token identity is not required in order to make sense of the causal role of the mental. I discuss Davidson's views on mental causation in Chapter 7, and I show there that instead of advantages, it has several problems. As concerns the realism issue, Davidsonian realism concerning the mental (as a type) is actually quite shallow. Since, on his view, an event is mental whenever it has a mental description, the reality of events *qua* mental is merely verbal; the mental is just a conceptual category.[38] I would emerge as a realist in this sense as well, although some would regard the ascription theory as an antirealist. After all, I accept that each time a mental state is properly ascribed to a subject, there are some physical events taking place in the subject's brain and his environment. But what more can be achieved by saying that this physical event is the very same mental event?

The next view that I am going to discuss is the identity of types. The type-type identity theory is compatible with a range of views on the nature of types. If one takes types as universals, then there is a correspondence between two pleonastic types and the natural basis would be a universal. If types are sets of tokens, then the type-type identity thesis implies a strong form of the token identity thesis, for in that case, each token from the *M* type is identical to a token from the *P* type. The token identity, as it is formulated above, does not require that the tokens with which *m* tokens are identified belong to the same physical type. To secure physicalism, it is enough for them to be physical.

There is one well-known problem with the type-type identity. In a nutshell, this is the idea that mental states can be multiply realised and hence they cannot be identified with any particular kind of realiser. This idea was originally presented by Putnam (1967, 1975b), and it is commonly regarded as dealing a decisive blow to type-type identity theory. Putnam's point was that mental states can be realised in different physical structures. This seems obvious in the case of different species. For example, pain can be ascribed to different organisms from a variety of species, but the chemical and physical structure or their nervous systems need not be identical. In some more far-fetched scenarios, the structure that might realise pain need not even be biological. Encouraged by the metaphor of computer software, which can be realised in a variety of machines, mental states were similarly taken to be features that could be specified independently from the hardware that happens to realise them. At this stage, the notion of realisation

38. "In my view the mental is not an ontological but a conceptual category. … To say of an event, for example an intentional action, that it is mental, is simply to say that we describe it in a certain vocabulary – and the mark of that vocabulary is semantic intentionality." (Davidson 1987b: 46.)

is not much more than a metaphor.[39] The same point can also be made in terms of ascribing the same state to creatures of different physical constitutions, having thus a different natural basis.

The argument against type-type identity theory is therefore the following: if in different organisms a given mental type can have a different natural basis, there is no justification for identifying this mental type with any single physical type (in the sense that both pleonastic types have the same natural basis). This would be chauvinistic with respect to other species, as it is sometimes put. Indeed, it would have the consequence that pain in animals, e.g., cats, is not the same kind as the pain that humans feel. Whatever the truth-value of such a claim, in any case, it was an unintended consequence of the type-type identity thesis. The considerations of multiple realisability showed that mental types cannot be identified with physical types.

There are ways of repairing the type-type identity theory. One could bite the bullet of chauvinism and still uphold the type-type identity, which is restricted to certain species. This would amount to a special form of functionalism, namely, realiser functionalism. Another option would be to reject type-type identity, while remaining faithful to the token-token identity theory. I have already discussed the token identity view and did not found it compelling. In the next section, I shall assess functionalism.

I conclude this part with a review of its main claims. I studied the notion of reduction by bridge-laws, which is still common in the philosophy of mind, and argued that it is not suitable for the purpose of ontological reduction. This led to the discussion of the strongest form of ontological reduction, namely the identity, which comes in the type and token versions. I tried to make sense of both in the framework of pleonastic and natural properties. From the deflationary point of view, the token identity theory does not seem to have any further positive content, once physicalism is presumed. To the type-type identity theory, I presented the common multiple realisation objection. My discussion of the multiple realisation was only preliminary. It paved the way to the discussion of functionalism and realisation that we embark on in the next section.

39. It is possible to talk about a system realising a Turing machine (Putnam 1975a), by which it is meant that the system is functionally isomorphic with a Turing machine (see also Kim 1996), but this notion differs from physical realisation or implementation, which does not require functional isomorphism.

2.5 Realisation and functionalism

The multiple realisation argument showed that a mental type like *being in pain* cannot be identified with any particular physical type, for it can be realised by various physical types. This raises the obvious question about the nature of mental types. There are two main functional conceptions of mental types which correspond to two kinds of functionalism: role and realiser functionalism. I discuss and criticise both of these in this section. My aim is to analyse functionalism from the ontological point of view. The ensuing discussion demonstrates that functionalism has its problems and that it may not be the best account of what it is to have a mind.

The common idea of functionalism, whatever its stripe, is that mental types have definitions in terms of causal or functional roles. It does not matter whether these definitions should be worked out a priori by reflecting on common usage or whether they should be based on the empirical results of scientific vocabulary.[40] In both versions, the claim concerning the availability of functional definitions is a claim about concepts or predicates; hence it involves conceptual reduction. In this step, it involves identifying pleonastic mental and functional properties at the same stratum. Here the 'is' of definition is used, as opposed to the 'is' of the identity theory, which involves the relation between the stratum of pleonastic properties and the stratum of the natural basis. The idea is thus that mental concepts are functional concepts, and, in the case of full-blown functionalism, it is assumed that the functional definition is exhaustive and complete: there are no aspects of mental states that are not captured by the net of causal-functional relations.

But what is this causal-functional role? It is assumed that there is an internal state which, on the input side, is causally related to the environment and, on the output side, yields certain behaviour. In addition to the relations between inputs and outputs, the relations between that state and other mental states also count in the specification of the role.[41] The functional role has thus three main components, all of which are specified relationally. To use a well-worn example, being in pain is being in a state that is caused by pricking, etc., which causes the subject to be in a state of distress, etc., and which leads the subject to say "Ouch!", or something similar. The emphasis on the conceptual reduction of mental predicate types to functional predicate types is reminiscent of analytic behaviourism, especially in its Rylean version. However, the crucial difference is functionalism's

40. Herein lies the difference between analytic functionalism (Armstrong 1968; Lewis 1972) and psychofunctionalism (Putnam 1967). The term 'psychofunctionalism' comes from Block (1978).

41. I ignore here the distinction between narrow and wide functionalism. They differ over the issue of whether external conditions enter into the specification of the functional role. This distinction does not affect the metaphysical point I am about to make.

emphasis on the existence of functionally described inner states. Ryle's (1949) approach provided analyses of mental concepts in terms of dispositions, which were spelled out by mere conditionals, for which no categorical basis in inner states was deemed necessary.

The functional definition of a mental state is but a first step in functionalism. No doubt, it is a controversial step. For example, there is strong disagreement over the question of whether qualitative states have functional definitions. There are also some worries about how the definition itself might proceed. I discuss the details of Lewis's functional analysis in Chapter 4. As for the present discussion, which is centred on the ontological issues, I waive the qualms about functionalisability, and see what follows if we proceed from the assumption that functional definitions of mental states are indeed available. As it turns out, even granting that much to functionalism does not make it compelling.

Suppose that we have an acceptable functional definition of a mental predicate. It is a complex predicate, specified in terms of relations between inputs, outputs, and other mental predicates. To avoid circularity, the mental terms would be replaced with quantified variables (see Lewis 1972, and Chapter 4), so that the specification itself would contain only relations between variables and other terms. The following step is finding an appropriate realiser in the organism (and in some views, partly in the environment). At this point, an ontological divide emerges in the functionalist camp. Sometimes this is not made explicit, but there are two stances occupied here. There is *role functionalism*, the view that functionally specified mental properties are role properties, which are ontologically separate from their realisers. And then there is *realiser functionalism*, which claims that mental states are just those realisers that meet the functional specification. I use the notion of realiser also in characterising the latter view, for using similar terms helps us to compare the views, and this usage is quite common. However, the notion of realisation itself is not perfectly clear, and the latter view does not really depend on it. It can also be called 'restricted type-type identity theory' or the 'functional specification view'.

The common assumption of these kinds of functionalism is that mental concepts *are* functional concepts, in the same sense that one would say that the concept of bachelor *is* the concept of an unmarried man. The disagreement, often implicit or sometimes mistakenly taken as merely verbal, lies at the stratum of inflationary properties. In a nutshell, this is the difference between saying that there exist both role and realiser, and saying that there exists only the realiser at the inflationary stratum.

I shall attempt to put the difference in terms of pleonastic and natural properties. Assuming that for both kinds of functionalism pleonastic mental properties are pleonastic functional properties (in the sense of definition and not identity),

we need not mention mental properties if the aim is to clarify the difference between these views. For simplicity, I also omit talk about pleonastic realiser properties. It is implicitly assumed that a natural realiser property has a physical description (which gives us a pleonastic physical realiser). I ignore the option, indicated by some functionalists, that a functional property can have a nonphysical realiser: some soul stuff, for example. If there are inflationary nonphysical properties, then presumably one should refrain from using the term 'natural' when talking about the natural basis.

A note on usage: a second-order role property is a property of having some realiser property. As such, it is a property of the property-bearer, not a property of a (first-order) property. Similarly, by 'role property', I do not mean 'property of a role', but 'property, defined in terms of functional roles'. The claims of the two types of functionalism would thus go as follows:

Role functionalism: For each pleonastic functional property-type, there is a natural second-order property-type, such that having any first-order natural property from the realiser set is having the natural second-order property.

Realiser functionalism: For each pleonastic functional property-type, there is a disjunction of natural species-specific realiser property-types, such that having any disjunct of the disjunction of such natural realiser property-types is having the pleonastic functional property.

To reiterate, having the pleonastic functional property is just another way of saying that one is describable as fulfilling a certain role. These formulations say that according to realiser functionalism, having a *natural realiser property* makes one being describable as fulfilling a role. According to role functionalism, by contrast, having a *natural role property* makes one being describable as fulfilling that role; the realiser comes into play only insofar as having a second-order role property requires having a realiser. (When I talk about having role properties, it is presumed that what one has are tokens of role property-types; the same applies for having realiser properties.) Note that both accounts would allow that there are pleonastic second-order properties. The difference is that there is no inflationary second-order functional type at the natural stratum according to the realiser view. According to role functionalism, the opposite is the case.

In what follows, I criticise both versions of functionalism. Let us begin with role functionalism. By assuming that there are natural role properties, role functionalism inflates the role description, apparently by adhering to the principle Φ, as characterised in Section 2.1. Such reasoning is guided by the assumption that whenever a predicate truly applies to an object, it does so in virtue of an inflationary property of that object and that every object describable by this predicate

must have the same property (cf. Heil 1999). But the principle Φ is not universally valid and it is not a good guide for compiling a list of natural properties. Heil (1999, 2003a) points out that a predicate can be truly applied to objects that are relevantly similar. Then the property is still applicable in virtue of some object's natural properties, and it is not required that all objects described by a given predicate have the same natural property. Moreover, one should leave open the option that some predicates apply to an object in virtue of an extended arrangement of natural properties both inside and outside the object. In such cases, it cannot be said, strictly speaking, that the applicability of a predicate is guaranteed by how the object is internally.

The inflation of role properties thus requires a supporting argument; it is not something that can be taken for granted. One might assume that the possibility of multiple realisation itself leads to inflationary role properties, but this is wrong. Multiple realisation shows only that mental types cannot be identified with (unrestricted) physical types. It does not follow that there are inflationary or natural mental types. This becomes evident when one notices that realiser functionalism is also compatible with the multiple realisation argument, but it does not presume the existence of special higher-order inflationary mental properties.

Inflationary role properties are metaphysically spurious. Why should we assume that when one is in a particular physical state or has a natural realiser property, one also has an additional *natural* or *inflationary* property, namely, the second-order property of having the first-order realiser property? I would not object if the second-order property were merely pleonastic. But when it is taken to be natural, then it is far from clear what it could be and how one could have it. If the natural role property is possessed in virtue of having a natural realiser, then what is the 'in virtue of' relation that instantiates a second-order property? Objects do not come to have inflationary second-order properties in virtue of the existential quantification over realiser properties that they have (cf. Kim 1998: 103). This may be the case for pleonastic entities, but not for the natural existents. For example, if natural properties are powers, then no one can seriously suppose that we add powers to objects just by quantification or by some other linguistic practice. Hence, the 'in virtue of' must be some sort of natural relation holding between lower-order and higher-order natural properties such that the latter do not reduce to the former. But what could it be? Presumably, it is the realisation relation.

But what is realisation? There is one common notion of realisation, stated, for example, in Bickle (2002). The idea is that property instantiation realises another property instantiation if it is necessarily sufficient for the latter. To put it more exactly, it says that x's having P realises x's having G just in case if x has P, x has G and for all x, it is necessary that if x has P, then x has G. Sometimes it is added that having P must somehow explain having G. However, this will not be enough

if one aims to understand realisation in the metaphysically more serious sense. As stated, the realisation relation is not in principle different from supervenience, and similarly, it is consistent with a perfect correlation of properties. The explanation requirement is too unspecific to save the day, for having *P* can explain having *G* without realising it. This is true whenever having *P* causes having *G*. The formulation above sheds no light upon the relation between having *P* and having *G*, and it fails to capture the idea that having *G* is a second-order property, which is had by having the realiser *P*. It can be added that *x*'s having *P* also constitutes its having *G*, but this will not do in general, since not all second-order properties, and mental properties especially, are constituted by first-order properties. The realisation relation is not to be analysed in terms of the part-whole composition either, as I shall argue shortly, since mental properties (or *x*'s having them) as functional roles are not parts of realiser properties (or *x*'s having them). I shall discuss these possibilities separately below, but the upshot is the same: when realisation comes down to constitution, then it becomes suspect in which sense the second-order properties are separate from their realisers.[42] Perhaps there is no good way to express this relation at all given the condition that the realiser property and the realised property should remain distinct. For it could be the case after all that from the ontological point of view, having a role property is not anything over and above having a natural property that fulfils that role.

Let us thus examine the idea that having a role property is constituted by having any instance of its realiser set and the role property is still distinct from its realisers. Thus far, I have not paid much attention to the fact that the specification of the functional role is causal. The role is specified in terms of the causal relations that the internal state has with other states and inputs and outputs. Perhaps this casts some light on the realisation relation or on the relation between roles and their realisers. Indeed, Pereboom and Kornblith (1991) have proposed a "token constitution" view, according to which tokens of higher-order mental properties are constituted by tokens of their realisers, while both the tokens and types of higher-order properties remain nevertheless non-identical with their bases. The view is expressed in terms of causal powers: "The causal powers of a token of kind

42. My rejection of higher-order inflationary properties does not lead me to deny that there can be different *layers* of natural kinds. For the issue of whether the higher-layer causal powers of natural macro-entities are excluded by the powers of lower-layer micro-entities, see Kim (1998) and Chapter 7. Note that Kim (1998) uses the term 'level' for the micro-macro relation, but this can be misleading, since there are also levels of explanation that need not involve part-whole relations. For example, personal level entities are not composed of parts of subpersonal entities. See also Hütteman and Papineau (2005), who draw a sharp distinction between the micro-macro or part-whole relationship and the relation of levels of vocabularies. For this reason, I use the term 'layer' to designate hierarchies of natural kinds.

F are constituted of the causal powers of a token of kind *G* just in case the token of kind *F* has the causal powers it does in virtue of its being constituted of a token kind *G*" (Pereboom and Kornblith 1991: 131).

The worry with such a view is this: why should we assume that if the causal powers of each role property token are had in virtue of the causal powers of property tokens that constitute role property tokens, then it does not follow that role property tokens are identical with their realiser tokens? It is not enough to stress the explanatory autonomy of higher-order properties, for explanation and generalisations can be underwritten by pleonastic role properties and the inflationary role properties are not required. The worry is metaphysical. For example, it is reflected in Kim (1998). He maintains the view that higher-order properties do not have separate causal powers, once it is conceded that they are had in virtue of the possession of their realisers. He subscribes to the causal inheritance principle: "if mental property *M* is realized in a system at *t* in virtue of physical realization base *P*, the causal powers of *this instance* of *M* are identical with the causal powers of *P*" (Kim 1992: 18). And if properties are individuated by their causal powers, then it follows from this principle that the instance of *M* must be identical with the instance of its realiser. In Chapter 7, I spend more time on the causal relevance of mental states. Here, the issue does not really concern how the mind can have causal powers; instead the worry is that role functionalism fails to make clear how can role properties remain distinct properties if one has them in virtue of having their realiser properties. All the tokens of role properties turn out to be identical with their realiser tokens. If one nevertheless insists that the role properties *qua* types must be conceived as separate from their particular realising tokens, and still inflationary, then, lacking any further reason, it is hard to understand this as anything other than an investment in the idea of inherently mental universals.[43] There is then a requirement to explicate the relation of 'in virtue of' that gives ground to talk about distinct inflationary role properties. Otherwise, role properties do not have genuinely distinct causal powers and turn out to be mysterious. This result should be especially upsetting for role functionalists, since they define roles in causal terms. In this sense, role functionalism has unwelcome consequences, and that constitutes a case against the account.[44]

---

43. In Armstrong's (1997) conception of universals, a universal is wholly present in each of its instances. Given physicalism, the instances of mental universals must be physical instances, but then it is hard to explain what makes them mental. The demand for inherently mental universals looks like a demand for dubious uninstantiated universals.

44. As will be seen in Chapter 7, the ascription theory need not presume that mental states are causal in the same sense that brain states are. According to the ascription theory, mental states have no distinct causal powers. In this regard, there is a similarity with role functionalism.

One such required explication is provided by Shoemaker (2001). However, as we shall see later, his view diverges from the token constitution account. He defends the idea that a role property has its own specific causal role that differs from the causal roles of its realisers (see e.g., Shoemaker 1981, where it is put in terms of causal powers). Heil (1999: 193) provides a critical discussion of this option and my treatment here is indebted to him. Imagine that there are two realisers, R_1, R_2, with various causal powers (P_1,…, P_5). However, there is a subset of causal powers (P_1,…, P_3) that they both share. Now, having a role property *M* is just having this subset of causal powers. In this view, having a realiser, e.g., R_1, is not necessary for having the higher-order property, since it can also be realised by R_2 or any other realiser that has suitable causal powers.[45] Still, having R_1 is sufficient for having the role property. The idea is that what justifies talking about the separate role property is the fact that since it has different causal powers than any of its realisers, it cannot be identical with any of them. Shoemaker constructs a new notion of realisation based on this idea. Assume that an object's having a property is this property bestowing causal powers to the object. Now, Shoemaker (2001: 78) says that property *R* realises property *M* just in case the powers bestowed to the object *o* by the realised property *M* are a subset of the powers bestowed to *o* by the realiser property *R*.[46] In case *M* is multiply realisable, its causal powers are a subset of the causal powers of each of its realisers.

There is a problem with this account, however. The assumption that one has the role property if one has either R_1 or R_2 does not give the role property distinct causal powers. The causal profile of each realiser exhausts all causal powers possessed by the state that fulfils the given role. As Heil (1999: 194) puts it, "the addition of *M* appears to require a kind of double counting of causal powers." It is not the case that the possession of the role property involves the possession of two sets of causal powers – one for the role, and one for the realiser. In each instance,

---

However, the difference is that in the case of ascription theory, this is an intended result and can be well accommodated in the whole framework, but in the case of role functionalism, it is a defect of the account when properties specified in terms of their causal roles turn out to have no separate causal powers. Hence the objection.

45. It is debatable whether it can be properly said that properties that are constituents of their realisers are higher- or second-order properties. If not, then Shoemaker's conception of realisation and his brand of functionalism differ from orthodox role functionalism. But even if a role property is understood in this way, it can still be said that it is a property that an object has in virtue of having some realiser property, and this makes it a higher-order property in my sense.

46. Shoemaker's formulation is given in terms of properties bestowing "conditional powers" to an object, thus presuming that the object has to have certain further properties in order to cause the required effect. As this has no influence on the present discussion, I omit this qualification.

there are only the powers bestowed by whatever realiser property there is. But then again, the question comes up: what more is there to having a higher-order role property than solely having the realiser? In either case, there just seems to be no room for two inflationary properties. Role properties are simply absorbed or exhausted by their realiser properties.

Shoemaker's suggestion that the powers of a role property form a subset of the causal powers of its realiser comes close to making the distinction between the role and realiser merely a matter of *describing* one set of causal powers as a subset of the other. The distinction between a set and its subset is relative to a description rather than independently constituted. If the aim is to defend the inflationary existence of a separate property, then this is not enough. If we come to count the distinct natural entities that operate in the world, we see that the subset of powers does not have any separate inflated existence from the total set of powers.

However, Shoemaker (2001) offers another reason for keeping role and realiser properties separate. His point is that the possession of one property *R* can include the possession of another property *M*, and the properties would still be non-identical when the property *M* is a constituent of *R*. One such case takes place when the properties stand in the *determinate-determinable* relation.[47] For example, an object can have the colour of scarlet, which is a determinate of the determinable property red. We may thus say that whenever an object is scarlet, it is also red, and that being scarlet includes being red, and still say that being scarlet and being red are not identical. On this point, it is important to pay attention to the difference between properties and their instances. Although Shoemaker (2001) holds the view that the powers that properties bestow to objects possessing them are essential to these properties, he no longer identifies properties with powers as he did earlier (see Shoemaker 1980). But he claims that the causal powers that *instances* of mental properties bestow on their subjects are subsets of powers bestowed by *instances* of realiser properties. Thus, he views the instantiation of a mental property as a constituent of the instantiation of its realiser property. "The instantiation of a realizer property entails, and might naturally be said to include as a part, the instantiation of the functional property realized." (Shoemaker 2001: 80–81.) Mental properties themselves, however, relate to realiser properties as determinables relate to determinates, being non-identical with them.

To this suggestion, the former objection about double counting natural entities still applies. In Shoemaker's view, mental properties are just those physical properties whose instances bestow on their subjects a subset of causal powers of all the instances of their realiser properties. If that is all there is to having a mental

47. Yablo (1992) has also argued that the relation between the mental and the physical is that of determination.

property, then, unless he accepts spurious inflationary mental types, Shoemaker's functionalism seems to succumb to a version of the token identity view, for in each instance of having a mental property, there is only one concrete instantiation of causal powers, not two separate instantiations. But apparently he aims to make sense of mental types, over and above their instances, being determinables of their realisers. However, the worry with this separation of properties and their instances is that its sole purpose seems to be to maintain the distinctness of functional properties, but this is *ad hoc*. Heil (2003c: 24) complains: "…it is not clear why we need properties anymore. Properties themselves, as distinct from and different in kind from their instances, seem to have no work to do: the action is all in the instances."

In addition, I would like to make two related observations. Firstly, I do not think that it is correct to view mental properties as determinables of more determinate realiser properties. In order for something to be a determinable of a determinate, both the determinable and the determinate must be properties in the same sense, differing only in respect of the determinacy of some aspect: in the present case, the determinacy of causal power. But mental properties and physical properties are not similar to that extent, if at all. It is misleading to think about the relation of mental and physical properties on the model of being red and being scarlet. Scarlet can indeed be viewed as a more determinate shade of red, but can we really make sense of a physical property being a more determinate aspect of a mental property?[48]

Finally, note that in Shoemaker's account, instances of mental properties are components of instances of realiser properties. Hence, physical property tokens are partly composed of mental property tokens. This is the opposite of the token-constitution view of Pereboom and Kornblith, which claims that mental property tokens are composed of physical property tokens. I am inclined to think that in the case of the mind-body relation, the idea of composition, in either direction, is a category-mistake. There is something deeply counter-intuitive in the notion that mental entities are components of physical entities. This encourages us to depict mental entities in a reified way, as things like tables, trees and neurones. This would be an inflation of mental properties to the point of its being a category-mistake. Obviously, the background idea is that mental entities are of the same kind as physical entities. But then the question is: what makes them mental? Why

48. Additional reasons for the claim that the determinate-determinable relation differs from the mental-physical relation are brought out by Ehring (1996). He specifies eight necessary conditions for the determination condition, and argues that in case of the relation between mental properties and their realiser properties, four such conditions are not satisfied.

were they singled out as mental in the first place?[49] If the "composition" idea is put in terms of causal powers, which obviously are not things, then it would still be misleading to imagine that mental entities have causal powers in the same sense that physical entities have such powers. The natural and naturalist way would be to think that there are only physical powers, and no inherently mental inflationary powers, rather than to assume that there are mental powers in the same sense that there are physical powers.

I end this discussion with an emphatic quote from Lynne Rudder Baker, who argues for the same point:

> Having certain kinds of brain states may be necessary for having beliefs; but it does not follow that particular brain states constitute particular beliefs. Brain states are ordinary spatiotemporal entities. Spatiotemporal entities are not widely scattered objects, but are compact objects that have more or less definite boundaries in space in time. A belief is no more constituted by an ordinary spatiotemporal entities than is the British Constitution. (Baker 1995: 184–185)

Given the objections to the separate existence of inflationary role properties, role functionalism seems to collapse into the token identity view.[50] Of course, it is possible for the functionalist to opt for *realiser functionalism*. But note that if a distinct inflationary higher-order role property is not to be had, then this has consequences for the antireductionist pretensions of functionalism. To be sure, the existence of Nagelian bridge-laws does not follow, but the damage is done to the looser sense of antireductionism, for mental properties do not seem to be over and above physical properties after all. Indeed, there is a sense in which that kind of functionalism, even after swallowing the multiple realisation argument, is still reductionist. In the case of realiser functionalism, where the requirement for separate inflationary role properties is dropped, the reductionist credentials become more evident.[51] Let us turn to an examination of this view. David Lewis (1966), and more recently, Kim (1998, 2005), have used functionalism to support type-type identities. There

49. See also Heil (2003c: 24), who worries that the idea of composition obliterates any difference between the mental and the physical and wonders what purpose is achieved by stressing that mental properties are realised by physical properties, but not reducible to them.

50. Carrier and Mittelstrass (1991: 61) express a similar contention about the Turing machine functionalism.

51. Block (1980) argues that the realiser functionalists are committed to role functionalism and hence they are mistaken to think that their view upholds physicalism. His discussion is potentially confusing, since he does not distinguish the two notions of a property. The fact that realiser-functionalists share the functional role specification of mental properties does not commit them to role functionalism.

are two main steps (three in Kim's case, but since we are here more interested in ontological than theoretical reduction, we do not have to discuss the third step: the construction of the explanatory theory, which explains how the realiser fulfils its role). The first step is taken in common with role functionalism: it consists in giving a functional analysis to a given mental concept. However, the specification of the role is crucially different. It does not say that having a mental property is having a separate second-order role property, but that it is having a realiser property, which satisfies the role-specification.[52] Still, the specification of mental property *M* is functional, pointing to the role (see e.g., Kim 2005: 111):

Having $M =_{\text{df}}$ having a property *P*, which plays the *C* role

The next step is empirical: the task is to find out which properties play the *C* role in a given species. According to Kim, this is the task of finding the realiser properties.

Here we have an explication of the notion of realisation as well. A realiser, according to Kim (1998: 20; 2003b: 579), is just any physical property that satisfies the role description *C*. This notion of realisation is better than that which was expressed in terms of nomological sufficiency. It is consistent with it, but it gives a solid metaphysical ground to the relation. If having *M* just is having a property *P*, which meets the *C* specification, then the connection between having *M* and having *P* is as tight as can be: it is identity. In this case, however, the respective identity statements are only nomologically necessary and the functionally defined designators can be called 'nomologically rigid' (cf. Kim 1998: 99; 2003b: 572–573). Note that this notion of realisation does not require that *M* and *P* are different and distinct *inflationary* properties. They are distinct as pleonastic properties, but out there in the world, in any given case, there is just one natural property-type and not two. In this regard, realisation functionalism comes close to the type-type identity theory. Of course, realiser functionalism is compatible only with the identity of types that is relativised to species or physical structures. We thus get identities between types that are species- or structures-specific. For example, pain felt by humans is identical with c-fibres firing; pain felt by dolphins is identical with some other neural state. The pain as such, as a general type, is not identified

52. Kim (1998: 98–99) wants to have it both ways. He takes mental property to be a second-order property of having a realiser property and then goes on to identify the property of having a realiser property *P* with the property *P* itself. However, a first-order property cannot be identical with a second-order property. Kim attempts to get around this problem by taking second-order properties as simply designators, or deflationary properties, but the problem still remains when one identifies pleonastic properties of different orders. The trouble can be avoided by defining a mental property as the property which meets a certain specification (Jackson 2002). In his reply, Kim (2002) seems to concede this point.

with any physical type. This is the lesson from the multiple realisation argument. Multiple realisation does not require that it has to be the same inflationary higher-order property that is realised in different structures. It is enough to talk about different natural properties satisfying a given description, and hence the different structures would only share a pleonastic property.

The point that realiser functionalism need not presume inflationary functional properties is also stressed by Kim (1998). He proposes to replace talk about inflationary higher-order functional properties (like the property of being in pain, minus its possible non-functionalisable component) with talk about mental functionalised designators. Another option would be to identify the higher-order functional type with the disjunction of its species-specific realisers. In that case, mental property-types would be reduced to disjunctive properties. It is also important to note that when Kim says that having *M* is having one realiser *P* or another, he is not compelled to accept the latter option, for this does not yet imply that *M* is identical with the whole disjunction of realiser types. Kim objects to inflationary disjunctive property-types, since they are not nomologically projectible. In order words, they cannot figure in laws.

It is somewhat unfortunate that the issue of the existence of natural disjunctive properties became thus connected with their suitability for figuring in laws (as in Antony 2003, and Block 1997). The background assumption is a kind of nomological approach to properties, which says that real properties are those that figure in scientific laws (cf. Mellor 1993). Another route to reach the same destination is to proceed from the nomological account of causation and to individuate properties in terms of their causal powers. No matter how the account may go, the endpoint is similar: there are no natural disjunctive properties. Instead of arguing over projectibility, I would prefer to phrase the argument against natural disjunctive properties in terms of the entities that ground the laws.[53] Then the claim would be that disjunctive property itself has no causal powers and it does not form a unitary natural kind.[54] The sole idea that there are disjunctive properties

53. If laws supervene on causal powers, laws might indeed be good guides for singling out natural properties, but I would object to views that regard laws themselves as more basic than properties or powers. Arguably laws and generalisations are pleonastic entities, having a place in explanation, but not necessarily in ontology as irreducible entities. Therefore, if the question concerning the existence of natural properties is put in terms of projectibility or suitability to participate in laws, there is a risk of confusing ontological matters with issues of explanation (cf. Jacob 2002: 651–652).

54. For criteria of being a natural kind, see Ellis (2002: 26–27) and Chapter 4. Even if laws of nature are used to explicate properties of natural kinds, the natural kinds as such are metaphysically more basic than laws.

comes from using an undifferentiated notion of a property and from the implicit acceptance of the principle Φ. Once it is made evident that there is no direct move from disjunctive predicates to disjunctive natural properties, support for the latter should wane. On this background, one can claim that the disjunction of natural property-types is not itself a natural property.

It is time to review where we stand at present. I have examined two versions of functionalism. The first and more common is role functionalism which identifies mental properties with special higher-order role properties. I rejected the existence of inflationary role properties on metaphysical grounds. It cannot be maintained both that role properties are realised by first-order properties and that they are distinct from the realisers. But if this is the case, then role functionalism collapses into token identity theory. There remains realiser functionalism, which in its most defensible version takes mental property-types to be pleonastic functional properties. In essence, it is a version of type-type identity theory, relativised to species or physical structures. In each case, at the natural stratum, we have a natural property that satisfies a certain specification. There are various species-specific type identities, but the overarching mental types as such are not identified with physical types.

What should we think of such a view? There are reasons to think that one cannot have even restricted type-identities. Even if we leave out mental properties that cannot be specified in terms of causal roles (and, for example, Kim (2005: 165) concedes that qualia are not functionalisable), there are additional problems even for those properties that are generally assumed to be functionalisable. Namely, given the plasticity of brains and individual differences, there is room for radical multiple realisability, which has the consequence that realisers differ also inside the species (see Block (1978) and Bickle (2003: 23) for a discussion of this worry). In the end, the realisations of functional roles might turn out to be strictly individual. In that case, the restricted type-identities are not to be had, and as in the case of role functionalism, all that remains is a token identity of each realised role-token with its particular realiser token property. On one reading, this is already implicit in my formulation of realiser functionalism. It says that that having any natural realiser property-type from the disjunction is having the pleonastic functional property. The relation of tokens is implied, if one assumes that there is no way to have a type-property without having a token property from this type. However, in such a case of radical multiple realisability, one cannot make sense of an identity relation between (restricted) natural types and functional types.

In later chapters, I shall present some additional qualms about functionalism that are concerned with the specification of the functional role, a topic that I have not touched on here. But for now, it is intriguing to observe the recurrent theme that runs through the various discussed accounts. Viewed from

the metaphysical perspective, the similar claim of the positions considered is that there is a certain internal natural basis that stands in relations of various strength to mental properties. As I noted in discussing token identity, I do not object to the general claim that in the case of every successful ascription, there are some processes going in the subject's brain. However, I object to attempts to construe this as a case of identity, a realisation relation or some other such fixed relationship. From the standpoint of the ascription theory, mental profiles can vary without any internal changes. For example, this happens when different mental states are ascribable in accordance with certain standards to internally identical subjects in the same situation, since the ascription-guiding external sources for these subjects differ. According to interpretivism, one may come to have different mental states irrespective of what is going on in one's brain. In this regard, brain states drop out of the picture: in everyday cases, they do not belong to the ascription sources and hence do not guide interpretation. The more relational and extrinsic the specification of a mental property, the farther away one inclines from the claim of identity.

2.6 Ascription theory and explanatory relations

Given nominalism about the mental and the rejection of inflationary mental entities, the issue of reduction comes to the fore in the ascriptionist approach only when it is viewed as an issue about the relation between vocabularies or theories. Therefore, I finish this chapter with a brief consideration of a position that an ascription theorist might adopt. This is not to suggest that these considerations follow a priori from the ascription theory itself. Rather, they elaborate an optional position that is consistent with my ascriptionist approach.

In general, I take mental states to be constructs that can be extracted from folk psychology. I do not assume that folk-psychological talk carries ontological commitments, and do not think that all folk-psychological categories have scientific value. On this backdrop, I advocate interpretivism, which aims to make sense of mental states in terms of their use in folk psychology, and which does not require that the ascription of mental states must be vindicated in the light of underlying physical states.

However, as concerns certain folk-psychological categories, I am, in fact, quite prepared to allow a certain restricted explanatory relation between pleonastic mental types and pleonastic physical types. Consistent with the idea that there are some brain processes going on in people who are subject to mental ascriptions, certain empirical regularities are to be expected. Namely, it is likely that certain folk-psychological distinctions provide grounds for looking

for certain structural and functional differences in brains. Presumably, common-sense notions should be idealised and, to use Quine's term, regimented, in order to be used in scientific theorising. In the end, this may give rise to some successor vocabulary of folk psychology, e.g., cognitive psychology that contains terms for which it is possible to point out an underlying physiological basis.[55] (This differs from the natural basis, which is a part of one's naturalist ontology that can correspond to certain pleonastic properties. The underlying basis in the brain is only a subset of that natural basis, which also includes the environment and other factors external to brains.) My point is that there are no principled reasons to deny that we can find such underpinnings that sustain certain regimented psychological constructs. Those underpinnings have always existed in brains; we just learn gradually that some distinctions we use in our everyday life echo the distinctions that can be found in brains. However, the full isomorphism of these two conceptual structures is unlikely, even if restricted only to a single capacity. In addition, it does not follow from this that neural theories would preserve or replace the normative relations between mental states. It is also important to note that given the assumption that it makes sense to point to certain brain processes as underpinnings of regimented constructs, this fact itself neither vindicates folk psychology nor makes its more isolated parts suspect. A good example is the case of memory. There is a rudimentary conception of remembering that we employ in our everyday practice. Since cognitive psychologists became interested in this subject, the concepts became sharpened, various distinctions were made, and quite specific neural mechanisms were found. In this respect, although underlying structures were found for the successor concepts of the ordinary notion of remembering, it did not follow that the remembering was reduced to those underpinnings or realised by them. At most, something to that effect can be said about the successor psychological construct, e.g., short-term memory. But it would be wrongheaded to say this about the unspecific folk concept of remembering, or about mental states in general as they are construed by folk psychology, given the conceptions of reduction and realisation discussed in this chapter. On the other hand, one might devise a more relaxed account of reduction than Nagelian and functional models. I leave open the prospect that, given some such account, those folk-psychological constructs, which are remodelled in cognitive psychology and

---

55. Cf. Dennett's remarks about the *intentional system theory*, which is idealised folk psychology, and *sub-personal cognitive psychology*, which deals with the implementations of the constructs of intentional system theory (see Dennett 1987: Chapter 3).

connected with certain brain processes, can then be said to be reduced.[56] Indeed, I would be happy to maintain reductionism concerning certain folk types, if *this* is all there is to reductionism.

The possibility of finding underlying structures does not undermine my point about the ascription-dependence of the mental, since I would argue that the mental is constitutively specified in terms of folk psychology, and that the everyday use of mental terms bears no specific commitments to internal states.[57] As stated, this leaves open the prospect that those terms can be further sharpened and that they would prove useful for scientific understanding. The resulting account would be an explanation of the processes that sustain certain functions that are specified in terms reminiscent of some older folk-psychological categories. This relation would be at the level of explanation, and thus differs from that of the realisation relation at least in two aspects. First, although the latter can include explanatory elements, as we have seen in this chapter, it is a metaphysical relation, which is basically a variation on the familiar theme of identity. Second, this relation would not involve the original pleonastic mental property, but the successor property.

Looking back, this chapter might seem rather tough going, but my task has not been simple. I aimed to provide indirect support for seeking an alternative account of the mind-body relation. For this purpose, I had to give some reasons why several widespread reductive approaches are not compelling. I attempted this by bringing out the commitments of different versions of identity theory and functionalism by viewing them through the lenses of my distinction between pleonastic entities and the natural basis. Then I provided a critique of these views, drawing in several places on the perceptive arguments of Jaegwon Kim and John Heil. Both the type and token identity theories were rejected and it turned out, perhaps surprisingly, that the commitments of role and realiser functionalism do not differ much from those of the already rejected token and restricted type identity theories. I also criticised supervenience theses and clarified the position of the ascription theory on this topic. It is consistent with the global supervenience of pleonastic mental entities on the natural world. Throughout the chapter, I urged caution against the inflation of mental properties, which leads to various metaphysically spurious positions. I also presented the main metaphysical views on the

56. There are some accounts of theory reduction that are more relaxed than the Nagelian account (Schaffner 1967; Hooker 1981), but they would not apply here since they involve the deduction of theories or isomorphism, and it is highly unlikely that theories of cognitive psychology could be deduced from neuroscience or even that these theories would necessarily preserve isomorphism of relations.

57. Furthermore, the possession of constructs of any successor psychological theory is also dependent on the ascription, but with sources different from common-sense ascription.

mind-body relation without assuming inflationary mental properties. The only exception that required inflationary role properties was role functionalism, which was rejected. Given the rejection of inflationary mental entities and their treatment as pleonastic entities, the mental does not pose special ontological problems and does not require metaphysical rescue in order to fit into the natural world. What remains is the conceptual task of clarifying the status of mental terms. I ended the chapter with a sketch of one feasible stance on the relation between folk psychology and neural explanation that is consistent with the ascription theory. Given the arguments and results of this chapter, it is necessary to investigate an alternative approach to the mental that attempts to make sense of it in the conceptual or pleonastic sphere and that does not picture minds and mental states as metaphysical entities. To this purpose, the following chapters are devoted.

CHAPTER 3

Interpretivism

3.1 Varieties of interpretivism

The common rationale of the views discussed in the previous chapter was to do justice to the idea that the possession of mental states constitutively involves the instantiation of some appropriate subset of the natural basis, namely and mostly, brain states, events or processes that either are identical with mental states, constitute them or satisfy the functional constraints of realising them. If any such account could work, then we could maintain a robust realism *cum* reductionism about the place of minds in nature. However, there are reasons, some of which were outlined in the previous chapter, for the suspicion that such approaches are not tenable as candidates for a complete and plausible account of the mind. In this chapter, I introduce and discuss an approach that downplays the alleged importance of brain states and other innards in coming to understand what it is to have a mind. This view is sometimes called 'interpretivism'.[1] Its basic tenet can be put quite vaguely to fit a variety of views under one heading by using a phrase from Dennett (1987: 342): there is "an element of interpretation" involved in determining what mental states one has. To be sure, interpretivism is not strictly opposed to the views discussed in the previous chapter. If taken as an epistemic or methodological stance, it can be coupled with a version of functionalism or identity theory. It can also be paired with some relaxed sort of behaviourism. And there is also an option of being an interpretivist about certain kinds of mental states (e.g., beliefs), while maintaining an identity theory about the rest (e.g., pains and sensations). Nevertheless, interpretivism has enough differences from the mainstream reductive naturalist position to be taken as a distinctive position, especially if it is understood as an constitutive account of what it is to have a mind. In this chapter,

1. This position bears slightly different names in the philosophy of mind and language literature. 'Interpretivism' is used by Fodor and Lepore (1992), Manning (2003) and Byrne (1998), who credits Mark Johnston's unpublished paper for the term. Some (e.g., Child 1994; Gauker 1988, 1994; Braddon-Mitchell and Jackson 1996; Heil 1998) use 'interpretationism'. Burwood, Gilbert and Lennon (1998) go even so far as to christen the view 'interpretationalism'. What is discussed under these headings is similar in broad outline, but it should go without saying that more specific formulations differ from one author to the next.

I discuss the work of Donald Davidson and Daniel Dennett. They have both advanced views that are commonly associated with the interpretivist label, and they are the main influences to the distinctive kind of interpretivism – the ascription theory – that I am outlining and defending in this book.

However, to say that someone's having a mental state, for example, John's believing that the cat has left the mat, involves an element of interpretation, is not perspicuous enough. This element may be required to specify the *content* of John's belief or it may be needed to specify the *state*, that is, whether it is a case of believing or guessing, for example; or it may be the case that the specification of both content and state requires interpretation. Another dimension that should be made clear is whether the element of interpretation is eventually dispensable or not. Let us say that interpretability is *constitutive* of mental phenomena in case its involvement in the explanation of this subject-matter is ineliminable. Otherwise, when interpretability may be involved but can ultimately be eliminated from the account of mental phenomena, it is *non-constitutive* (Child 1994: 48 develops a similar usage). In addition, the locution that interpretation is constitutive of mental phenomena requires further elucidation as to whether it is meant that interpretation is constitutive of the explanation of the *nature* of belief and meaning or it pertains only to the *possession* of belief, whereas the nature of belief – understood as an answer to the question 'What is belief?' – is to be given independently of interpretivist considerations.[2] Hence, some clarifications are in order.

Let us start from the latter issue and postpone the question of whether interpretability should apply to mental contents or mental states. If we proceed from the distinction between explaining the *nature* of a mental state (with a given content) and explaining the *possession* of a mental state (with a given content), then the interpretivist position with respect to the former task would claim that there is no understanding of what a given kind of mental state is without an understanding of how and why mental states of this kind are attributed. In other terms, this would be a view that the very concept of a mental state (or concepts of some mental kinds like the concept of belief, the concept of desire, etc.) incorporates a reference to interpretability.[3] It should be obvious from this formulation that one

2. For brevity and stylistic reasons, I sometimes write about interpretivism about belief or thought. However, unless stated otherwise, I intend that these claims apply also to interpretivist positions about other mental phenomena.

3. Gauker (1988, 1994) develops an interpretivist account about the nature of propositional attitudes. In his view, "an account of what beliefs are need not be anything over and above an account of their attribution" (1994: 293) and hence he frames his account of the nature of beliefs and desires entirely in terms of what it is to attribute beliefs and desires. I will take issue with his views in Chapter 4.

may hold this brand of interpretivism about some classes of mental states, e.g., propositional attitudes, while rejecting it for other classes. On the other hand, interpretivism with respect to the *possession* of mental states would be the view that a complete account of what it is to have a mental state of some kind, e.g., belief, involves or is exhausted by an account of attributing mental states of this kind. The ascription theory is interpretivist about the possession and not about the nature of mental states, for reasons that are explicated later. The same is true of Dennett's account, which can be more readily construed as an interpretivism about the possession of mental patterns than the nature of mental states. Matters are less clear with Davidson, since there are ways of reading his holistic approach so that, according to him, it is impossible to explain what belief is without explaining its role in relation to other mental states, objects of belief and interpretative triangulation. However, since he seems to be insensitive to the distinction between the nature and possession of mental states, he applies interpretivism largely also to the possession of mental states. As the issue of possession is common to Dennett's and Davidson's accounts, I will concentrate on interpretivism mainly as a view about the possession of mental states.

The question of whether interpretability is constitutive or non-constitutive is related to the question of whether interpretability is necessary or sufficient for mental phenomena. Given the formulation of the contrast between the constitutive and the non-constitutive, if interpretability is a necessary condition for having mental phenomena, it follows that it is also ineliminable from the correct account about the possession of mental phenomena and is therefore constitutive. But if interpretability is only sufficient but not necessary, it follows that interpretability is non-constitutive of mental phenomena. For if interpretability is a sufficient but not a necessary condition for having a belief, then one could also have beliefs in some other way and the reference to interpretability is not required in an account of what it is to have a belief. Then again, there can be other construals in which the 'constitutive account' is understood as one that gives necessary and sufficient conditions. However, to say that interpretability is constitutive of thought only in case it is both necessary and sufficient for thought is to declare that only quite a radical form of interpretivism is an interpretivism proper and to dismiss some moderate forms of interpretivism as non-constitutive accounts. The formulation from which I proceed, by contrast, allows us to distinguish between claims that interpretability is constitutive of having thought and that interpretability exhausts having thought. Note that in this way, constitutive vs. non-constitutive and necessary vs. sufficient would emerge as different distinctions. The reason is that the issue concerning the necessity or sufficiency of interpretability depends on how rich one's notion of interpretability is, while the constitutivity issue does not (cf. Child 1994: 40). Consider two positions. One takes interpretability as a mere

necessary but not a sufficient condition of belief, which has to be complemented by additional conditions. The other position includes such conditions already in the concept of interpretation and claims that interpretability, thus understood, is both necessary and sufficient for belief. In a sense, there is only a verbal difference between these views. However, a non-constitutive account cannot be updated to a constitutive account simply by adding conditions to the notion of interpretability. After all, the crucial issue that divides them is whether the very notion of interpretability is eliminable from the complete account of belief.

Let us clarify the topics of interpreting content and interpreting states. The question as to whether there is an element of interpretation only in the specification of content or only in the specification of states remains an issue in both kinds of interpretivism. However, the question itself is of secondary interest. For it seems to be the case that there is no way to ascribe pure content without also ascribing some sort of attitude to the content. It can be a bare "colourless state" at the very least, but even this is a state. Of course, it is possible to abstract content away from its state and to make claims only about the former, but the possession of content as it is thought or entertained cannot be separated from the state of its being thought or entertained.[4] Note that this leaves open the possibility of having bare states without contents. Hence, this claim does not amount to a full intentionalism, which states that all mental phenomena have intentional content (cf. Crane 2001: 8). Now, if it is true that content is always had in conjunction with some state, then this is a good reason for taking the interpretivist line to apply both to contents and to states. Indeed, once one has taken the interpretivist approach, it would be hard to make sense of the idea that only the content-specification involves interpretation, whereas states wear their names on their sleeves. And the idea that only the states need interpretation, whereas contents float free of interpretation, is downright unintelligible. Therefore, the putative separation of contents from states yields no additional support to either brand of interpretivism.

In view of the interpretivist project of explicating the possession of mental phenomena, the aforementioned considerations support the point that the possession of content is to be explicated in conjunction with the possession of its adjacent state and that both are subject to interpretation. Hence, when I write about the possession of content or what it is to have a mental state, it is usually assumed that the claims apply to the possession of a mental state with a particular content.

4. One can self-ascribe contents that do not involve the specification of the state under which this content is apprehended, but this is different from the issue of possession. In the latter case, respective states and contents go together.

There is a potential line of argument from the alleged fact that mental content always has some specification (be it either *de re* or *de dicto*), and from the premise that every specification is interpretable, to the claim that interpretability is part of an account of propositional attitudes. However, this is too weak to support the stronger interpretivist claim that the nature of propositional attitudes is to be explicated in terms of their interpretation or attribution, and this is what an interpretivism about the nature of the mental tries to establish.

As a matter of clarification, it should be noted that there is another project which can also be pursued under the heading of explaining the nature of content. This is explaining aboutness or reference: the explication of what makes it the case that the content *p* is about *P*'s. An interpretivist answer to this question is not forthcoming (see Chapter 1). Even our exemplary interpretivists – Dennett and Davidson – do not espouse a fully interpretivist answer in this respect, although their proposals need not conflict with the interpretivist framework.[5]

Thus far, I have marked out the version of interpretivism that is relevant to my project as an interpretivism at the level of possession. I have identified this position in a somewhat cautious manner in terms of the involvement of an interpretive element in the possession of mental states. It is time to spell out different ways in which this can be understood.

Interpretivism can be presented as a claim that interpretability is involved in the *explanation* of the possession of mental phenomena, or as a claim that interpretability is a condition of the very *possession* itself. The distinction here is that between epistemology and metaphysics – the distinction between what there is and how we know what there is. Occasionally interpretivists move from one claim to another without paying proper attention to this change and, hence, there is little wonder that a common accusation is that interpretivism

5. Dennett flirts with the teleofunctional account of content: "My theory of content is functionalist: all attributions of content are founded on an appreciation of the *functional roles* of the items in question in the biological economy of the organism" (Dennett 1998: 359). But in his view, functional roles depend on the interpretation. Davidson (e.g., 2001d) can be read as expressing support for a mixture of causal and inferential theory of content, which says that the content of concepts is partly determined by their causal relations with environmental objects in the triangular situation of concept-learning and partly by their relations to other concepts. See also his externalism-supporting discussion of Swampman's words lacking content, since they are not learned in a proper context in Davidson (1987a: 443–444). These statements fit in with the general picture of interpretivism if read as claiming that the correct interpretation should be guided by the objects in the subject's environment and ascribing holistic complexes of thoughts. An interpretivist should not assume that causal and inferential relations or teleofunctional roles confer content to states and words somehow brutely, bearing no constitutive connection to interpretability.

incorporates a confusion of epistemology and metaphysics (see Child 1994: 23). However, it is likely that such moves are due to the fact that interpretivists incorporate epistemic elements into their account of the mental. For if having *x* itself is an epistemic matter, then it is enough to phrase the analysis of having *x* solely in epistemic terms. Moreover, given this metaphysics of having *x*, a correct account of what it is to have *x* must involve mention of epistemic matters. In other words, interpretivists make a metaphysical claim about mental phenomena that their possession is epistemically constrained. Obviously, this claim needs an argument. Some considerations for this are given as we proceed. Here, the point is just that interpretivists are generally not guilty of a simple confusion of what there is and how one knows about it.

This discussion also provides us with a distinction between the epistemic or methodological and the metaphysical reading of interpretivism. As a *methodological* account, interpretivism diminishes just to the bundle of maxims and principles that would help us to interpret one another. Taken on the *epistemic* reading, it is compatible with the case in which the possession of mental states is manifestable in such a way that whenever one has a mental state, there is an interpreter, possible or actual, who can find out which mental state one has. On this reading, it comes down to an optimistic claim about the epistemic circumstance we are in with respect to the mental. Neither of these versions implies anything about the metaphysical constitution of mental phenomena or their possession. For imagine a mechanism that is an absolutely reliable detector of 'quans', which are just physical events of a certain type. In this case, it is true that all the mechanism registers about the emergence of quans is all there is to discover about them. Given the reliability of the mechanism, this may well be the best method for learning about quans. However, it does not follow that quans or their emergence are constituted by the readings of the mechanism. By contrast, as a *metaphysical* claim, interpretivism concerns the constitution of the possession of mental states. In its strongest version, it comes down to the claim that my believing that such and such is exhausted by me being interpretable as believing such and such.

This distinction connects with another distinction that can be exploited here. Consider the following biconditional:

> ($\forall x$) (*x* has a mental state *s* with content *p* if and only if the mental state *s* with content *p* is ascribable to *x* according to a certain nontrivial criterion.)

For present purposes, let us take ascribability as a type of interpretability. Whereas 'interpretation' is a broader term, 'ascription' can be taken in the sense of a specific assignment of properties to a subject. I have used the biconditional in this formulation to bring out the intended distinction. But it should be noted that an ordinary truth-functional biconditional may not be suitable for expressing all

versions of interpretivism, the ascription theory among them. One reason for this is that it fails to express the direction of determination. The biconditional is true, whenever the propositions on the left-hand side and on the right-hand side are both true (or both false), but it does not tell us, for example, that the proposition on the left-hand side is true *in virtue of* the truth of the proposition on the right-hand side. Let us read the biconditional from the right-hand side to the left-hand side. The idea is then that a certain kind of ascription *determines* whether a subject has a mental state of some kind.[6] If read from left to right, by contrast, it suggests that the ascription merely *reveals* the subject's already existent mental states via some special mechanism, and this is consistent with the epistemic reading mentioned above. Giving priority to different sides of the biconditional constitutes the "Euthyphro Contrast", discussed at length in Wright (1992: 108ff). Take the statement:

> For all *x*, *x* is loved by the gods if and only if *x* is pious.

The Euthyphro Contrast is a contrast between two claims. '*X* is pious, because it is loved by the gods' gives a projectivist reading, where the piety is had in virtue of being loved by the gods, while '*x* is loved by the gods, because it is pious' gives a reading where *x* is pious independently from the love of the gods (cf. also Child 1994: 48).

In light of the aforementioned distinction, interpretivist positions form an array in which the role of interpretation changes from metaphysical determination to epistemic revelation. At one extreme, we have a position that can be called *pure ascriptivism*: one's having mental states is exhausted by one's being accordingly interpreted or interpretable.[7] Interpretability is thus a necessary and a sufficient condition for having mental states. Pure ascriptivism is a constitutive account of mental-state possession. It is ineliminable from that account, since while it supposedly exhausts the condition of having mental states, there is no way to explicate this condition without referring to interpretability.

---

6. This coincides with the common reading of biconditionals as reductive, on which the right-hand side is taken to give a reductive analysis of the left-hand side.

7. The term comes from Clark (1993: 214). In his portrayal, this is the view that "being a genuine believer is, in a certain sense, essentially a matter of how others might find it profitable to treat you". He attributes this position to Daniel Dennett. Arguably this is not Dennett's position, however, at least not in *The Intentional Stance*, to which Clark refers. Indeed, Dennett (1987: 15) considers an account he calls "interpretationism", which "likens the question of whether a person has a particular belief to the question of whether a person is immoral, or has style, or talent, or would make a good wife." Given that the latter are a mere "matter of interpretation", whereas in Dennett's view "belief is a perfectly objective phenomenon", he rejects this account because of its inherent relativism.

A less extreme position is *intermediate interpretivism*. Its idea is that the ascription of mental states uncovers an intentional pattern, which is not merely a matter of projection of interpreters. However, the interpretability is ineliminable from the full account of the pattern. This approach faces the task of explaining in what sense the pattern is there, while still being dependent on interpretation. The ascription theory is an example of interpretivism of this variety. The notions of pleonastic property and pleonastic fact enable us to account for the "fragile" sort of reality of the intentional pattern. Namely, there are intentional properties and facts only in the pleonastic sense, and the possession of such intentional properties and the holding of such intentional facts are to be explained by reference to their ascription conditions. On the one hand, the ascription of intentional patterns is not answerable to independently existing intentional facts; on the other hand, by being constrained by available data and principles of interpretation, it is not arbitrary. The ascription does not track the very facts ascribed; given other relevant facts, it non-randomly determines these facts. The existence of such facts is no more mysterious than the cases of sentence's being true and predicate's being correctly applied. These claims need not be confined to intentional properties. In my view, this can be said about all mental properties and facts. I thus endorse interpretivism about all mental phenomena, whereas Dennett and Davidson seem to treat mainly intentional states from the interpretivist point of view.

Finally, the other end of the spectrum is occupied by *revelationism*.[8] It postulates that the existence of intentional patterns is independent from interpretation. It can still be regarded as a version of interpretivism, if the intentionality of patterns is explicated by reference to interpretation. However, the interpreter or the interpretability is not constitutive for this account. The interpretivist element may have been included to ease the exposition, or to make an epistemic point about the interpreter's access to intentional patterns, but ultimately it is dispensable.[9] In the first case, it would be interpretivism only in the methodological sense. In the second case, it seems to presume some miracle of revelation, for there seems no reasonable ground for the interpreter's epistemic privilege in respect of intentional

---

8. In general, revelationism is the view that a certain kind of contact with a property reveals its essence. This view is sometimes held in respect of experiences – having an experience is being in a position to know all of its properties – or colours – the colour experience reveals the essence of colours. For such a usage, see Johnston (1992).

9. For example, David Lewis utilises the construct of an omniscient interpreter in order to clarify the supervenience of intentional facts on physical facts. His question – "how do *the facts* determine these facts?" – makes no commitment to the ineliminability of the all-knowing interpreter (Lewis 1974: 333).

patterns, especially if it is assumed that intentional facts like ordinary physical facts are metaphysically independent from the observers.

As already noted, I opt for intermediate interpretivism. Pure ascriptivism makes having a mind too dependent on the whim of the interpreters. There seems no place for objective constraints on the interpretation in this approach. Revelationism is either mysterious or not a proper version of interpretivism in any stronger sense, if it turns out that interpretability is not an integral part of the account.

Much has been written on the topic of the reducibility of intentional patterns. It is important to note that this issue differs from the question of the centrality of interpretability. For example, Dennett (1987) has argued that intentional patterns are such that they can be discerned only in intentional vocabulary, from the "intentional stance", as he expresses it. This differs from the claim that the discernibility of intentional patterns is constitutively bound up with interpretation. It is one thing to say that an intentional vocabulary is necessary for discerning intentional patterns; it is another thing to say that intentional patterns must be interpretable. It follows that one cannot conclude merely on the basis of Dennett's defence of the stance-dependent reality of intentional patterns that he subscribes to intermediate interpretivism in my sense. These two claims can be run together only in case the interpreter is an ineliminable component of the intentional stance. (I will return to this point later in Section 3.3).

There is no conclusive verdict on Davidson's position vis-à-vis various interpretivist options as outlined above, but it is safe to assume that he rejects pure ascriptivism and that he differs from intermediate interpretivism also. Since he holds the identity theory about mental tokens, claiming that each mental event is identical to some physical event, he can be interpretivist only about mental types. However, in view of the fact that he repudiates properties (understood as types) in his ontology, his interpretivism can only be epistemic or methodological (at least, if taken in his own terms). He assumes that "an observer can under favorable circumstances *tell* what beliefs, desires and intentions an agent has" (Davidson 1982: 321, my italics). His use of the word 'tell' supports the revelationist reading. On this reading, the interpretation only reveals an agent's intentional profile and is not constitutive of it. Elsewhere he has said explicitly:

> Our justifying evidence for attributions of attitudes to others is their behavior, sometimes observed, but more often correctly assumed or inferred from context. Our attributions *do not constitute or create* the attitudes. There are endless attitudes people keep to themselves and even when they don't we are apt to get them wrong. (Davidson 2001b: 299, my italics)

The following picture therefore seems to emerge: an interpreter can in certain epistemic conditions come to ascribe intentional states, but their existence in each particular case is constituted by particular physical events.[10] However, there remains a problem in this view. If, as is the case for Davidson, an event's being mental is just a matter of its falling under a mental description, then there is a need to explain what determines the application of mental predicates. It cannot be determined by a mere occurrence of an event, having a physical description, for there are no laws linking the application of physical and mental predicates. And if the applicability of mental predicates is not constituted by interpretability either, then we lack the required explanation. One way to clarify Davidson's assertions is to point to the ambiguity of the notion of constitution. In the sense used in this chapter, '*x* is constituted by *y*' means that *y* (is part of what) determines the nature of *x*, by making it a thing of the kind it is.[11] But constitution can also be understood in the sense of material composition. In this sense, to say that '*x* is constituted by *y*' is to make a claim about the material make-up of *x*.[12] Given this distinction, there is a potential position such that, although physical events constitute or compose mental events, they do not constitute or determine their nature as mental, which instead is determined along interpretivist lines. It is subject to dispute whether such a position is open to Davidson. I will return to this issue at the beginning of the next section.

A degree of disquiet may arise from my talk of pleonastic properties being constituted by other properties. For, one might enquire, how can pleonastic properties, being properties only in the deflationary sense, have enough substance to be constituted by something else? Part of this concern may result from not paying heed to a distinction between constitution and composition. It is true that only inflationary entities can be composed of other entities. But in order to give a constitutive account of an entity, it need not be inflated. A constitutive account of a pleonastic property is a theory about this particular concept from which a partic-

---

10. However, in a later footnote, Davidson declines to give primacy to the physical, opting for a neutral monism, which allows that "the same *things* are described now as physical, again as mental or biological, etc." (Davidson 2001a: 281–282, fn 1.)

11. The addition of the 'is part of' is meant to leave open the question of whether *y* is a necessary and a sufficient condition for *x*, or is merely necessary.

12. The idea that belief tokens are constituted by brain states in the sense of composition is discussed and rejected by Lynne Rudder Baker (1995). In her view, the statement that beliefs are composed of brain states is not meaningful and amounts to a kind of category mistake. (Baker 1995: 188). Indeed, in view of the role of beliefs in folk psychology, it makes little sense to think of particular instances of having beliefs as being composed of neural events, in the sense that mounds are composed of pebbles. See also my remarks in Section 2.5.

ular pleonastic property is obtained by the something-from-nothing transformation. If the constitutive account concerns the *nature* of a pleonastic property, then it may also be viewed as an account of the meaning of the respective predicate. If it is a constitutive account of the *possession conditions* of a pleonastic property, it concerns the application conditions of a predicate. In my view, as I explained in Chapter 1, the latter account is not exhausted by the former account, at least in the case of mental phenomena. This is constituted by interpretability in the sense (already explicated) that the explanation of why a given mental predicate applies to an agent refers to interpretability.

Having thus introduced the basic ideas of interpretivism as well as the variety of ways in which these ideas can be cashed out, I devote the remaining parts of this chapter to a close study of the interpretivist ideas of Davidson and Dennett, which bear important connections to the ascription theory that is largely provoked and inspired by them. In the next section, I summarise Davidson's main points about interpretation together with some critical remarks. Then I give an exposition of Dennett's intentional stance theory. Neither of these tasks is straightforward, for both philosophers have changed their views more or less over the years. I have attempted to interpret them charitably and to support my attributions with a wealth of references and quotes. Some objections to Dennett's and Davidson's accounts and some ways of countering them are discussed as well. Finally, I consider some sources of divergence from their approaches that have influenced the eventual form of the ascription theory.

3.2 Donald Davidson and radical interpretation

Davidson is often taken as an advocate of interpretivism. In the previous section, I presented the conjecture that his interpretivism is of the methodological and epistemic kind and that it is not presented by him as a metaphysically constitutive account of thought. I return briefly to the exegetical issue of whether he holds constitutive interpretivism about mental types (as opposed to particular mental events). Let us begin with two quotes. The first is a classic one giving credence to the claim that Davidson is indeed an interpretivist, and the other supports the epistemic reading.

> What a fully informed interpreter could learn about what a speaker means is all there is to learn; the same goes for what the speaker believes.
>
> (Davidson 1983: 433)

Here we have a clear statement of the idea that beliefs and meaning are fully accessible from the viewpoint of an interpreter.

> Anomalous monism does not suggest that mental events and states are merely projected by the attributor onto an agent; on the contrary, if holds that mental events are as real as physical events, being identical with them, and attributions of states are as objective. (Davidson 1997: 112)

Anomalous monism is Davidson's version of token-token identity theory. Its monistic aspect was discussed in the previous chapter. It is the idea that mental events are identical with physical events, and that their mentality consists only in their having mental descriptions. However, Davidson calls his monism anomalous to highlight his contention that there are no strict laws connecting mental and physical descriptions. As for the quotation, then, pure ascriptivism is rejected there, and the reality of mental events is underlined. At least in this quote, there is insufficient material to support the view that Davidson sees his interpretivism in any stronger sense than the epistemic. It is also clear that, in his view, no such interpretivism follows from anomalous monism.

Note that we need to distinguish between the question of the reality of mental states and the issue of whether interpretability is constitutive of the mental. These are different questions, and it should be noted that it need not follow from constitutive interpretability that mental states are not real. It might seem as if there is a choice between only two options: either mental states are real and objective or they are a mere projection on the part of the interpreter. But this overlooks the option of intermediate interpretivism. However, even if one does not pay attention to this distinction or assumes that there is no stable intermediate option, there is still a difference between the reality of a particular mental event and the reality of mental types. If there are mental events, identical with physical events, then obviously there is no question of their reality. But it is still questionable to what extent they owe their reality to their mentality, or in other words, to the fact that they fall under mental types. In Davidson's view, an event is mental if it has a mental description. There is thus a task of explaining why they merit their mental descriptions and what is involved in applying mental descriptions to events, and, at this point, there is even room for constitutive interpretivism to creep in. Indeed, this would be a constitutive interpretivism concerning mental types in the first place, but since each mental event involves a particular case of applying mental predicates, one can argue that interpretative considerations must also play a role in determining the character of particular mental events as mental. Of course, this need not imply that Davidson himself considered this line of thought.[13]

---

13. Given my notion of metaphysics, which allows for pleonastic mental entities, such an interpretivist account about mental types could well be called 'metaphysically constitutive', but only if one assumes that this is all that can be said about the mental, metaphysically speaking.

In what follows, I shall devote some space to explicating Davidson's views on interpretation, which play a central role in his philosophy of language. This is important for the purposes of the present book, since his work contains several ideas that are relevant to understanding how interpretation can take place and which constraints it should follow. The ascription theory that will be developed in Chapter 5 differs from Davidson's account in various ways. Some of these differences may be due to a difference in emphasis only, while some are more principled and require an argument. I point out the differences as we go along.

Davidson's views on the interpretation of propositional attitudes should be uncovered in the meaning-theoretical context. His aim was not to give an account of belief *per se*, but to say what is involved in the understanding of sentences. In explaining the latter, Davidson has gradually developed an integrated account of mind and meaning. What became known as the 'Davidsonian project' is an attempt to provide a semantics for a natural language on the basis of the Tarskian account of the truth-predicate. This theory was to provide a finite basis for the implicit knowledge of speakers and outline how we can learn an infinite number of sentences on the basis of finite elements. A condition of adequacy for a theory of meaning is thus that it must be compositional, allowing that the truth conditions of a sentence derive from the semantic properties of its parts. Another condition is that the theory must be extensional, that is, that it should not rely on intensional constructs like 'means that' (cf. Miller 1998: 247). Davidson (1967b) argued that we can get hold of such a theory by utilising Tarski's convention (T). The resulting theory of meaning would thus produce for every sentence of a particular language L a theorem of the following form:

(T) s is true_L if and only if p.

Here, the structural description of the sentence in language L goes in place of '*s*' and '*p*' is replaced by the very sentence itself (given that the meta-language contains the object-language) or otherwise by the meta-language translation of *s*. What is pivotal is that such a theorem gives the truth condition of *s* in L.[14] Thus we get (to use the classic example):

'Snow is white' is $\text{true}_{\text{English}}$ if and only if snow is white

Davidson's anomalous monism apparently commits him to saying that mental events have more substantial metaphysics.

14. Note that L is not a variable, but part of the truth-predicate and, therefore, Tarski's scheme does not give a general account of truth-in-L, for any L. Cf. Davidson (1990b: 285).

Davidson's proposal was that the truth condition of a sentence can be taken as a specification of its meaning. However, it would be misleading to say simply that in Davidson's account the meaning of a sentence is its truth condition (cf. Lepore and Ludwig 2005: 166, n. 136). His account is rather more complex. Namely, the idea is that a theory of meaning for a language can be offered in the form of a theory that generates for every sentence of a language a theorem giving its truth conditions. One way to approach the problem of meaning is to ask what it is to understand or know the meaning of utterances. Hence, in some places, Davidson has expressed the task of providing a theory of meaning in terms of explaining what is involved in the understanding of the speech of another speaker: "A theory of truth for a speaker is a theory of meaning in this sense, that explicit knowledge of the theory would suffice for understanding the utterances of that speaker" (Davidson 1990b: 312). There is a substantial issue concerning how to choose the theory that generates T-theorems and what the applicable constraints are. At this stage, the considerations of interpretation become relevant to the choice of the required theory. I shall discuss this issue in the following sections.

Finding a theory that gives appropriate truth conditions is not a straightforward matter. It may seem that in order to assign truth conditions to a sentence, one should already understand that sentence. For otherwise one can assign as a truth condition of *s* any such true (extensionally equivalent) sentence that preserves the extensional structure of a theory, while the result is clearly unacceptable. For example

'Snow is white' is $\text{true}_{\text{English}}$ if and only if grass is green

Since the sentence that replaces *p* is true, the theorem above is extensionally correct. However, it is clear that this result fails to interpret 'Snow is white'. It may seem that the condition of extensional correctness is too coarse to yield a satisfactory assignment of meaning.

Davidson cannot solve this problem by adding the requirement that the sentence in place of *p* in the scheme, 'grass is green' in the present example, should be the correct translation of *s*, for that would make his account circular (for this argument, see Miller 1998: 264). After all, a common way of unpacking the notion of a correct translation is to say that the correct translation is the one that preserves meaning. Hence, one cannot explicate the notion of the meaning of a sentence by saying that it is given by such truth conditions that provide meaning-preserving translations. In sum, Davidson was thus faced with two complications. The first is the question of how to constrain the truth theory so that the resulting biconditionals are both extensionally correct and interpretive. The second is the threat of circularity: how to avoid using concepts that presuppose the notion of meaning in elucidating that very notion.

Davidson worked around these problems in the following way: he stipulated that only a theorem that follows from the theory that provides us with an acceptable interpretation of a speaker gives the meaning of a sentence. To provide an acceptable interpretation of a speaker's utterances, the truth theory should satisfy certain interpretative constraints and it should be empirical – or fit with the evidence available to the interpreter. In order to counter the circularity threat, the truth theory, or interpretation theory as it is sometimes called, should be supported by evidence that can be specified in non-semantic terms, or, at least without presuming the notions of meaning and translation (Davidson 1974: 311–312). This point merits further discussion later in the chapter. In order to rule out inappropriate T-theorems, the interpretative constraints are applied holistically at the level of the whole theory. It is assumed that the interpretative constraints lead us to reject any truth theory that yields inappropriate T-theorems.

Partly influenced by Quine's (1960) device of radical translation, Davidson articulates the project of interpreting a speaker without presupposing knowledge of the meaning of his utterances as a "radical interpretation". Apart from certain similarities, there are also several differences between Davidson's and Quine's approaches. This is not the place for a prolonged discussion of these, but some deserve to be mentioned (cf. e.g., Evnine 1991: 99–101 or Davidson 1979). Quine's radical translation results in a mapping between the sentence of two languages. This need not uncover the meaning of the sentences, but Davidson's aim is to explicate what would suffice for knowing the meaning of the sentences of the object language. Hence, if one knows the truth theory resulting from the interpretation, one understands the object language to which the theory is supposed to apply. Another difference concerns the evidence available to the interpreter. For Quine, the evidence for radical translation was purely behavioural: namely, the native speaker's patterns of assent to and dissent from sentences. Davidson does not attempt to reduce claims about meaning to behavioural evidence; as will be argued shortly, his requirement is more modest, trying to show how notions like meaning and belief can be extracted from concepts that are still intensional, but can be more closely related to the behaviour. There are some commentators who assume that the only evidence available to a Davidsonian radical interpreter is observable behaviour (e.g., Rawling 2003: 93). Lepore and Ludwig (2005) also argue that it is Davidson's view that, as a matter of convenience, the radical interpreter can rely on facts about the attitude of holding true, but these are to be "grounded ultimately on more primitive behavioral evidence" (Lepore and Ludwig 2005: 158). However, they give no hints about how to effect such grounding and the quotes they offer in support of this claim fail to establish conclusively that this is indeed Davidson's considered view.

As already mentioned, Davidson's position is that knowing the truth theory for a language and knowing that it conforms to certain formal and empirical constraints suffices for understanding the speaker of that language (see Davidson 1973a: 326–327). The empirical constraint is that the truth theory should be confirmable without presuming the facts the theory is supposed to explain. In effect, this is the standpoint of the radical interpreter. Hence, the connection between a theory of meaning and interpretation can be put in the following way: only a T-theorem that follows from the truth theory that fits with the empirical evidence accessible to the radical interpreter gives the meaning of a sentence. The interpretation is called 'radical' because it does not assume any prior knowledge of a speaker's beliefs and the meanings of his utterances. The evidence that is available to the radical interpreter

> must be of a sort that would be available to someone who does not already know how to interpret utterances the theory is designed to cover: it must be evidence that can be stated without essential use of such linguistic concepts as meaning, interpretation, synonymy, and the like. (Davidson 1973a: 316)

The evidence that a radical interpreter can rely on includes information about the speaker's environment and his non-intentionally described behaviour, in terms that do not imply anything about his beliefs and the meanings of his utterances. Most importantly, however, Davidson (1973a: 322) proposes that the radical interpreter could know what kinds of sentences the speaker *holds true* at a given time. That is, the radical interpreter is allowed to know the speaker's attitudes towards the truth-values of his sentences.[15] Although the attitude of holding true is a kind of belief, and thus seems to be excluded from the evidential base of the radical interpreter, Davidson notes that this attitude can be identified without individuating its content. The idea is that the interpreter can know that a speaker holds true sentences s, s_1 and s_2 while having no clue what these sentences mean: "he may know that a person intends to express a truth in uttering sentence without having

15. Mainly in his (1980b) essay and in some other papers, Davidson has extended his basic account of interpretation with the aim of incorporating degrees of belief and strengths of desire. Instead of the attitude of holding true, the attitude of "preferring one sentence true rather than another" is taken as primitive, which is a factor in an agent's meanings, relative desirabilities about the truth of sentences and subjective probabilities. Assuming that the logical constants can be identified from the patterns of preferences, Davidson draws on the Richard Jeffrey's (1983) version of decision theory with some modifications in order to assign conditional expectations of utility and subjective probabilities to sentences. Then, as in the previous model, the principle of charity is used to specify the truth conditions of sentences. The central assumptions of this model, especially those related to interpretivism, do not differ from his account as discussed here. For a thorough comparison of these two models, see Laurier (1999).

any idea *what* truth." (Davidson 1973a: 322.) However, even if knowing that one holds a sentence true need not entail knowing the meaning of this sentence, it cannot be denied that we are dealing here with an intentional attitude, and not with a behavioural event under purely physical description (cf. Evnine 1991: 100; Heal 1997: 184; Miller 1998: 268). Behaviour can indeed be physically described as a sequence of movements of a physical body, but there is no physical description of an event of holding something true in the offing. For that would presume that there is a physicalist or non-semantic account of truth available, which does not seem to be the case, and this runs contrary to Davidson's own inclination to take truth on board as a basic, unanalysed notion. But then it cannot be said that all the evidence available to the radical interpreter is specified in non-semantic and non-linguistic terms, as Davidson (1974: 311–312) notes. However, his more careful claim, quoted above, that the evidence must be stated "without essential use" of linguistic concepts would still remain standing.[16]

This invites the following considerations. Given that the starting point of radical interpretation is not wholly non-semantic, Davidson's project would then seem to be more modest, articulating certain connections that hold among the central concepts of intentional discourse. In outlining Ramsey's version of decision theory, which is the structural prototype of his own model, Davidson (1980b: 4) notes that decision theory does not reduce intensional concepts and says that it is still "an important step in the direction of reducing complex and relatively theoretical intensional concepts to intensional concepts that in application are closer to publicly observable behavior." It is apparent from the context of this paper that his own approach to meaning and belief merits the same characterization. There is no objection to such an approach, but it casts doubt on Davidson's answer to the circularity objection. For it was the desire to avoid circularity that seemed to

16. In the context of discussing his later project of extracting a unified theory of meaning and action on the basis of the subject's preferences with respect to the truth of his sentences, Davidson (1984a: 17) writes: "The idea has been that the pattern of uninterpreted responses of a certain sort is enough to allow the assignment of meanings to words, the attribution of beliefs, and the determination of objects of desire. The responses are uninterpreted only in that the ordinary propositional contents are not known; the basic data for interpretation are *still intensional in character*, since they involve the concepts of desiring sentences to be true. The relation between a person and a sentence he desires to be true is purely extensional, since it holds however the person and the sentence are described. This does not suggest that the words expressing the relation can somehow be reduced to purely extensional concepts; I'm sure they cannot be." (Italics mine.) This concedes again that the attitudes of holding or desiring true are intensional. It is correct to say that a person would desire a sentence to be true no matter how the sentence is described, if it means something like 'I prefer the truth of a sentence written on top of this page', but it is incorrect if one means that one's preferences will always remain the same if the content of the sentence is described in coreferential terms.

motivate constraining the truth theory by requiring that it be assigned from the austere position of the radical interpreter. But if he also allows semantic evidence, then there is no good answer to the circularity objection. Perhaps it is indeed the case that there is no way to illuminate semantic notions like meaning and translation in purely non-semantic terms, and that our only option is to explain "thicker" semantic notions in terms of "thinner" semantic terms like holding true. Such a difference, however, if it can be clearly specified at all, is a matter of degree, and it would provide no principled reason to link the confirmation of T-theorems to the standpoint of the radical interpreter. If the aim is to find out which T-theorems are interpretive, then why assume that this has to be found out from the position of the radical interpreter? Once the requirement that the theory must be based on non-semantic evidence is weakened, there is no impediment to confirming T-theorems from the standpoint of the ordinary interpreter, whose evidential basis is richer, allowing some background knowledge about the language and other ascribable mental states of the subject. Of course, in order to be informative, the evidential basis should not contain information about the meaning of the very sentences one is interpreting or facts about the very mental states that are being ascribed.

One answer that Davidson might give to this objection is that the evidence available to the radical interpreter is the *sufficient* evidence required for confirming T-theorems or interpreting a speaker. "The point of the 'epistemic position' of the radical interpreter is not that it exhausts the evidence available to an actual interpreter, but that it arguably provides sufficient evidence for interpretation." (Davidson 1994a: 121.) However, from the fact that certain conditions are in principle sufficient for some *x*, it does not follow that these conditions do in fact or ordinarily or mostly lead to *x*. Taken in this way, interest in his project of conceptual explication would diminish, if one aims to keep closer contact with the application of mental concepts in their common-sense context. The ascription theory is guided by the assumption that we can find out something constitutive about mental states if we pay attention to what is involved in the usage of mental concepts to make sense of oneself and others. It would be more reasonable for this project to consider *all* such typically available evidence that can be invoked in the interpretation rather than to limit oneself to what would suffice for interpreting in the artificial circumstances of a radical interpreter.

In Davidson's approach, radical interpretation relies on a restricted range of evidence, but at the same time, quite high demands seem to be made of the interpreter's abilities. The ascription theory attempts to find a balance between these extremes. Besides removing the restriction on the evidential base, no presumption is made that the ascriber would possess the knowledge of theorems yielded by the Tarskian truth-theory or grasp the principles of some decision theory. The

reason for this is that if the possession of mental states is constitutively connected to their ascribability, then the realisation of the ascribability conditions should not fall beyond the ken of ordinary ascribers, or be attainable only by some ideal agent. Davidson may protest at this juncture that the apparatus he is explicating describes rather the competence of an interpreter and there is no need to assume that anyone actually knows all this explicitly. For example, he has made the following clarification:

> In any case, claims about what would constitute a satisfactory theory are not … claims about the propositional knowledge of an interpreter, nor are they claims about the details of the inner workings of some part of the brain. They are rather claims about what must be said to give a satisfactory description of the competence of the interpreter. (Davidson 1986: 163)

Nevertheless, there is no reason to assume that the description of the competence of a radical interpreter illuminates the competence of the non-radical ascriber who attempts to settle which mental states someone has. On the one hand, it has not been shown that the competence of such an ascriber must be described in truth-theoretic and decision-theoretic terms. On the other hand, it is clear that since such interpreters rely on a richer evidential base, this must also be part of their competence.

In sum, given that Davidson allows the radical interpreter to access information about the attitudes of holding true, the distinction between the radical and the ordinary interpreter becomes blurred. There seems no principled reason as to why one should not enlarge the evidential base of the interpreter. If the purpose is to explicate what the minimal base is from which the interpretation *could* proceed, then it becomes clear that Davidson's project differs from the project that I am presently developing. This does not entail, of course, that his project is devoid of intrinsic interest or even that it is not instructive to examine it in the present book. In the following discussions, we shall see that Davidson's ideas about indeterminacy and charity are relevant for the ascription theory as well.

Let us proceed with Davidson's account of interpretation. The evidence about what sentences speakers hold true (in a language L) is given in the following form (cf. Davidson 1973a: 322–323; Miller 1998: 268):

> (HT) John holds $true_L$ 'It is snowing' at time *t* if and only if it is snowing near John at time *t*.

Based on such evidence, the interpreter is seeking to confirm sentences in the form:

> (TC) S's utterance 'It is snowing' at time *t* is $true_L$ if and only if it is snowing near S at *t*.

If the relation between (HT) and (TC) were simply a matter of generalisation, then in order to confirm or to raise the likelihood of (TC), one would need to investigate whether such facts as those stated in (HT) are widespread in the L language community.[17] However, as Davidson (1974: 310–311) claims, the fact that a speaker holds true a sentence *s* depends on both what he means by *s* and what he believes. If all the interpreter knows is (HT), he is in no position to determine what John means, when uttering: 'It is snowing', if he does not know what John believes. For John might believe that the weather is dry and intend to express this by uttering 'It is snowing', or he might attempt to mislead the interpreter. Conversely, there is no direct link between the attitude of holding true and beliefs. For example, it can be the case that John holds 'It is snowing' true, but since he means by this that it is raining, one cannot infer that he believes that it is snowing. What is needed is a wedge to break into the interdependent sphere of belief and meaning. Davidson's idea is to rely on certain assumptions about a speaker's beliefs in order to fix the meanings of their utterances. For this purpose, he makes use of the *Principle of Charity* as a constraint of interpretation.[18] The idea of charity is to assume that speakers' beliefs about what is going on around them in plausible proximity are mainly true. The principle obliges one to assign such "truth conditions to alien sentences that make native speakers right when plausibly possible, according, of course, to our own view of what is right." (Davidson 1973a: 324.) However, Davidson has expressed the principle of charity in different terms in various places, and it is by no means obvious that all his formulations are equivalent. Sometimes the principle is expressed in terms of maximising the agreement or minimising the disagreement between the speaker and the interpreter: "Torn between the need to make sense of a speaker's words and the need to make sense of the pattern of his

17. Some philosophers (e.g., Hacking 1986) have worried about the potential conflict between Davidson's project of giving a truth-theory for a language and his later insistence that interpretation does not require a prior shared language and that there is "no such thing as a language" if language is something that consists of a vocabulary and a set of conventional rules shared in the language community (Davidson 1986). A way to reconcile this is to note that the interpretation is given only to an individual speaker and the notion of language community is constructed on the basis of individual speakers that can all be subjected to the same interpretation theory (cf. Lepore and Ludwig 2005: 290).

18. The term "principle of charity" originates from Wilson (1959). It was also invoked by Quine (1960: 59) as a principle of radical translation. For Quine, the principle consisted in the claim that "assertions startlingly false on the face of them are likely to turn on hidden differences of language". In explaining its motivation, he said that "one's interlocutor's silliness, beyond a certain point, is less likely than bad translation", meaning that if one's translation makes the speaker assent to logical inconsistencies or falsehoods, then it is likely that the translation is mistaken and should be rejected.

beliefs, the best we can do is choose a theory of translation that maximizes agreement." (Davidson 1968–69: 138.) If, in addition, it is assumed that the interpreter's own beliefs are true, then the difference between maximising true beliefs and maximising agreement does not loom large. Davidson (1975: 21) also seems to assume that maximising the similarities among the beliefs of the interpreter and the speaker minimises the chance that both of them have largely erroneous beliefs about the world. Nevertheless, the formulation of the principle in terms of agreement between the interpreter and the speaker is secondary in relation to the formulation in terms of truth, and the latter is more central to Davidson's project of confirming a truth theory on the basis of facts about what sentences the speaker holds true.[19] What is required is a principle that allows one to move from (HT) to (TC), to move from S's holding $true_L$ *s* to *s* being $true_L$. This task requires the truth of S's beliefs, for this helps to fix the link between the obtaining of a circumstance *p* and S's having beliefs about the circumstance, which in turn allows us to deduce what S means with *s*. Since the initial stages of radical interpretation concern basic observational beliefs, the sentences the speaker is holding true concerning his immediate surroundings, it suffices to formulate the principle of charity that initially constrains the radical interpretation in the following way:

> The Principle of Charity: *Ceteris paribus*, a speaker's beliefs about the conditions in his proximity that play a causal role in triggering his beliefs are true.[20]

This is not Davidson's own wording, but I hope that this more specific formulation performs better the role the principle is required to play in the first stages of radical interpretation.[21] On the basis of charity, it thus becomes possible to

19. In some places, Davidson stresses that what is at stake here is the truth merely by the interpreter's lights; in other places, he ventures to claim that what seems true to interpreters must in most cases be true *simpliciter*. For our purposes, the more modest claim is sufficient.

20. Note that this formulation need not commit one to a causal theory of belief and thus push one away from interpretivism, since the idea that beliefs are caused by environmental conditions can itself be part of the (folk-psychological) theory that is applied in interpretation. If charity is a condition for interpretation, then similarly, it would be a condition of interpretation in folk-psychological terms that one takes the central folk-psychological assumptions on board.

21. The interpretation of Davidson's views on the principle of charity is a complex matter. For a detailed study, see Lepore and Ludwig (2005: Chapters 12–14). They suggest that a modified "principle of grace" would suit Davidson's purposes even better. This principle places a condition on a speaker which says that "if a belief of his is prompted by the environment, then it is about a prompting condition" (2005: 194, n. 169). This allows for a straightforward individuation of the content of belief, once the prompting condition is identified. For further studies on principles of interpretation, as well as modifications of the principle of charity, see e.g., Grandy (1973), Lewis (1974), Gauker (1986), Henderson (1987).

infer T-theorems that make up a truth theory for a language. The next step is to ascribe contents to further attitudes and truth conditions to further utterances, not directly related to changes in the environment, while checking the ascriptions in light of overall coherence and behavioural predictions. In later stages of interpretation, further principles may come to play a role. A central constraint is to assume that the speaker's beliefs and thoughts are consistent. Davidson has explicitly distinguished the principles that require ascribing true beliefs and consistent beliefs:

> The Principle of Coherence prompts the interpreter to discover a degree of logical consistency in the thought of the speaker; the Principle of Correspondence prompts the interpreter to take the speaker to be responding to the same features of the world that he (the interpreter) would be responding under similar circumstances. Both principles can be (and have been) called principles of charity: one principle endows the speaker with a modicum of logical truth, the other endows him with a degree of true belief about the world. Successful interpretation necessarily invests the person interpreted with basic rationality.
> (Davidson 1991: 158)

The principle of correspondence conforms to what I have specified above as the principle of charity. Davidson views both principles as constitutive conditions of the possibility of interpretation. The correspondence or truth requirement is motivated as the only means for breaking into the interdependent circle of belief and meaning, whereas the coherence principle comes from certain a priori assumptions about the nature of agency. The idea is that a rational agent must obey the canons of logic and decision theory.[22] These canons are normative in the sense that they prescribe how a rational agent is supposed to reason. Through the principle of coherence, these canons also guide interpretation. Davidson offers an example of how the norms of reasoning influence interpretation:

> If someone believes Tahiti is east of Honolulu, then she should believe Honolulu is west of Tahiti. For this very reason, if we are certain she believes Honolulu is west of Tahiti, it is probably a mistake to interpret something she says as showing she also believes Tahiti is west of Honolulu. (Davidson 1990a: 24)

---

22. Here is one characteristic passage: "If we believe certain things, logic tells us there are other things we ought or ought not to believe at the same time; decision theory gives us an idea of how beliefs and values of a rational man must be related to each other; the principles of probability theory sets limits to how we may rationally adjust our faith in a hypothesis given that we accept certain evidence; and so forth." (Davidson 1990a: 24.)

There is no clear indication in Davidson's work of what the exact measure of consistency that is minimally required from an agent is.[23] In addition, the rules that determine what kind of thinking counts as consistent could originate from various theories and areas. One could require the *logical* consistency of beliefs, as well as coherence in the sense of *probabilistic* coherence or in some other informal sense. Davidson apparently means all of the foregoing. He notes that the principles come from decision theory and then continues with the list: "the basic principles of logic, the principle of total evidence for inductive reasoning, or the analogous principle of continence. These are the principles shared by all creatures that have propositional attitudes." (Davidson 1985a: 351.)[24] Here, Davidson is propounding a constitutive claim that these principles determine who counts as a rational agent, that is, who those individuals are to whom we can apply concepts of propositional attitudes. This allows him to argue further that what deviates from these principles is not interpretable. However, Davidson does not think that we could ever settle on a fixed inventory of all such rationality principles. Rationality is "uncodifiable", as it is sometimes put (see Child 1994: 58ff). This is one of Davidson's reasons for why the mental cannot be reduced to the physical (assuming the model of theoretical reduction), for there cannot be strict biconditional laws relating mental ascriptions and physical conditions. In addition, if the rationality principles are not clearly circumscribed, then it seems that the question of whether someone is interpretable or not does not have a clear-cut answer either. This issue is also related to the indeterminacy of interpretation that will be discussed shortly.

Sometimes Davidson adds an extra twist and says that what is not interpretable is not a thought at all:

> The issue is not whether we all agree on exactly what the norms of rationality are; the point is rather that we all have such norms, and that we cannot recognize as thought phenomena that are too far out of line. Better say: what is too far out of line is not thought. It is only when we can see a creature (or 'object') as largely rational by our own lights that we can intelligibly ascribe thoughts to it at all, or explain its behaviour by reference to its ends and convictions. (Davidson 1990a: 24)

23. Contrast this with Christopher Cherniak's (1986) notion of the minimally rational agent, requiring as a descriptive condition that such an agent generally eliminates *some* inconsistencies among his beliefs and must be able to make *some* sound inferences from his activated belief set.

24. The *principle of continence* commands one to "perform the action judged best on the basis of all available relevant reasons" and the *principle of total evidence for inductive reasoning* instructs one to give one's "credence to the hypothesis supported by all available relevant evidence" (Davidson 1969: 112).

At first glance, this quote seems to imply that *interpretability* is a condition for thought or that what *we cannot recognise* as rational is not thinking. But Davidson then seems to backtrack immediately and says that what departs from the *norms of rationality* too much, and hence fails to show the required amount of coherence, is not thought. This seems to imply that norms of rationality themselves, rather than whether we can recognise them in operation, are what really matters. And this leads again to the question discussed in Section 3.1, as to whether interpretability is eliminable from Davidson's account. Let us distinguish three options. On one reading, interpretation is just a device of explication that can be ultimately eliminated from the explanation of what constitutes being a thinker. This account therefore states that the constitutive conditions are merely that one's thought follows the patterns of rationality, is suitably connected with the environment, etc. However, another reading points out that since coherence and other norms are necessarily imputed only in the process of interpretation, what it is for a pattern of thoughts to be coherent and rational cannot be explicated without the requirement that the pattern must be interpretable. The third option would be to deny the primacy of either rationality and coherence or interpretation: for only a coherent pattern of thought is interpretable and once someone is singled out as a subject of interpretation, there is a presumption that a coherent pattern is to be ascribed. In fact, there is not enough textual evidence to yield conclusive support any particular reading.[25] In view of Davidson's later work that stresses the holistic interdependence of central intentional notions like belief, truth, and objectivity, one can conjecture that the last option conforms best to his outlook, but given the considerations brought out earlier in this chapter, the first, epistemic reading is also plausible.

Davidson has often stressed that the assumptions of truth and rationality are not optional guidelines, but are essential in interpreting and constituting speakers and agents. See e.g., Davidson (1973–74: 19) or the following passage:

> The policy of rational accommodation or charity in interpretation is not a policy in the sense of being one among many possible successful policies. It is the only policy available if we want to understand other people. So instead of calling it a policy, we might do better to think of it as a way of expressing the fact that creatures with thoughts, values, and speech must be rational creatures, are necessarily

25. In describing what is involved in interpreting speech, Davidson (1986) acknowledges that a theory of meaning is not enough for interpreting a speaker. What is required in addition is the decision to apply that theory in a given case, and the knowledge underlying this decision is not part of the theory of meaning: it cannot be "regularized" in the form of finite rules, for it is "just about everything we know" (Davidson 1998: 112). This seems to contain the seeds of an argument for the ineliminability of an interpreter, but it is too underdeveloped to yield any conclusive verdict on Davidson's position.

> inhabitants of the same objective world as ourselves, and necessarily share their leading values with us. We should not think of this as some sort of lucky accident, but as something built into the concepts of belief, desire, and meaning. (Davidson 1984a: 18–19)

The upshot of proceeding from these assumptions is that people are seen as predominantly rational and holding true beliefs. This has invited a considerable amount of criticism, for it is well known that ordinary speakers and agents fail to meet the high normative standards of rationality in various ingenious ways. The question of how much rationality is required in interpretation is discussed further in Section 3.4.1. For now, it has to be pointed out that Davidson does not deny that agents can exhibit irrationality and can have mistaken beliefs. However, he argues that the attribution of particular irrational beliefs or occasional mistakes would still presume that the agent has otherwise largely rational beliefs. This is due to the holistic nature of interpretation. A single belief can be ascribed to an agent only by ascribing to him several other beliefs and propositional attitudes as well. What the interpreter ascribes to his subject is thus a whole interlocked web of beliefs and meanings and sometimes this can include various irrationalities, provided that their ascription is warranted.

The principles of interpretation do not yield a unique assignment of truth conditions to a speaker's sentences and propositional attitudes to a speaker. There can be different, nonequivalent patterns of meaning and belief that are all equally ascribable to an agent, even if all the constraints are considered. Davidson (1973b: 720) admits "cases where all possible evidence leaves open a choice between attributing to a speaker a standard meaning and an idiosyncratic pattern of belief, or a deviant meaning and a sober opinion." That the ascriptions of beliefs and meanings can balance each other is for Davidson the main reason for the *indeterminacy* of the mental. But Davidson differs from both Quine and Dennett in that he does not draw sceptical conclusions from this about the nature of meaning or about the reality of belief. Both Quine (1981: 23) and Dennett (1991b: 49) claim that in the case of different, but equally acceptable schemes of translation or interpretation, there is a genuine indeterminacy, since there are no intentional facts to which the correct translation or interpretation scheme corresponds. Thus, the indeterminacy should be distinguished from the mere underdeterminacy of theories by the data. In the latter case, different theories would fit with the available evidence equally well, but unlike the case of indeterminacy, there would still be a matter of fact as to which theory is the correct one (Quine 1987: 10). This is the sceptical twist: there are no facts of the matter about what the subject believes or what his utterances mean. As long as the interpretations fit with the behavioural evidence, they are acceptable as determinants of a subject's mental profile, and no further mental facts are forthcoming that can play a part in determining the correctness of the interpretation.

By contrast, Davidson regards indeterminacy as the harmless thesis that "one can capture exactly the same attitudes in systematically different ways." (Davidson 2001b: 300.) There is no assumption that due to such indeterminacy, talk in terms of mental events or meanings is somehow second-rate or nonfactual.[26] Of course, Davidson does not countenance mental entities in the inflationary sense, but in his view, the question of whether there are mental entities plays no role in the issue of indeterminacy. What matters instead is whether ascriptions of mental states can be objectively true or false. Davidson (1997: 121) grants that the "grounds for choosing among conflicting hypotheses" concerning mental ascriptions are intersubjective, but in his view, this kind of objectivity is enough when the ascriptions of propositional attitudes are at issue. When there is indeterminacy, it ensues from using different vocabularies to describe the same fact. He draws an analogy with different systems of measurement for the same phenomenon. It is possible for us to use either Fahrenheit or Centigrade scales to measure temperature. If we have an algorithm for converting from one scale to another, it can be said that both scales measure the same phenomenon. Or as Davidson (1998: 99) puts it: "What matters is what is *invariant*. … Only invariances are 'facts of the matter.'" Similarly, different interpretation theories can be constructed that fit with the available behavioural evidence and with evidence concerning the sentences the speaker holds true. The indeterminacy that Davidson allows is thus indeterminacy pertaining to the choice of interpretation theory. However, once we have fixed the scheme of interpretation, there is no internal indeterminacy in the assignment of truth conditions and beliefs.

What are the "facts of the matter" invariant in different interpretation theories? This is a matter of some controversy. One option is that it is the relations among ascribed propositional attitudes and the meanings attributed to sentences that are invariant.

> Each of us can think of his own sentences (or their contents) as like the numbers; they have multiple relations to one another and to the world. Keeping these relations the same, we can match up our sentences with those of a speaker, and with the attitudes of that speaker, in different ways without changing our minds about what the speaker thinks and means. …our sentences can be used in endless different ways to keep track of the attitudes of others, and of the meanings of their sentences. (Davidson 1998: 99–100)

26. "Indeterminacy of meaning or translation does not represent a failure to capture significant distinctions; it marks the fact that certain apparent distinctions are not significant. If there is indeterminacy, it is because when all the evidence is in, alternative ways of stating the facts remain open. … Indeterminacy of this kind cannot be of genuine concern." (Davidson 1974: 321–322.)

This is to say that we can use different but convertible sentences of our own language to provide the truth conditions of a speaker's utterances. But this kind of benign indeterminacy differs from what Davidson has brought out as the main source of indeterminacy – namely, the possibility of adjusting a speaker's beliefs while keeping the meaning and ascribing different meaning while keeping the beliefs fixed.[27] For imagine someone uttering 'This sunset is beautiful' while observing the sunrise. On one interpretation, we may ascribe the usual meaning to her sentence with a mistaken belief that it is evening and it is time for dusk. Another interpretation would be that she has mixed up the words or means sunrise by 'sunset', and that she has the true belief that it is dawn. Clearly, these interpretations ascribe different structures to the speaker, and moving from one interpretation to another would constitute a 'change of mind' as to what she believes or means.

In fact, Davidson claims later on the same page something that seems incompatible with the required way of understanding what is invariant. In discussing Quine's scepticism concerning facts of meaning, he notes that Quine's scepticism was directed at the idea that each word has one real meaning and says that "this negative point does not entail that there are no facts of the matter; *the facts are the empirical relations between a speaker, her sentences, and the environment*. This pattern is invariant." (Davidson 1998: 100, my italics.) It is difficult to reconcile this claim with his other remarks on the topic. The empirical relations between the speaker, her sentences, and the environment are comprised in the sentences the speaker holds true in a given situation. This is the data from which the radical interpretation proceeds and to which the resulting truth theory or interpretation theory must conform. Hence, these facts cannot be the attributed facts that are presumably invariant in different notational variants of the interpretation theory. Another problem is that Davidson wants to maintain the realist attitude concerning the mental in contrast to the sceptical stance of Quine and Dennett. Therefore, he has to locate the invariants to the "intentional realm". However, the empirical relations between the speaker and the environment are neither intentional nor mental relations. Both Quine and Dennett would agree that interpretations should match behavioural facts, and they can well allow that relations between the speaker, her sentences, and her environment belong to the set of behavioural evidence. Hence, if this is what makes indeterminacy innocuous in Davidson's view and underlines the reality

27. "When all the evidence is in, there will remain, as Quine has emphasized, the trade-offs between the beliefs we attribute to a speaker and the interpretations we give his words." (Davidson 1973a: 327.)

of the mental, then his contention is unwarranted and he has not provided a serious alternative to Quinean scepticism concerning intentional facts.[28]

In sum, there are two options concerning what the invariant nodes in the interpretation schemes can be. According to the first option, the indeterminacy is a matter of different mappings of the interpreter's sentences and subject's attitudes and sentences. The problem with this account is that different mappings can entail attributing different intentional profiles to the subject that are not mere notational variants. The second option is to take the relations between the subject and her environment as invariant nodes, but this fails to support Davidson's insistence that the indeterminacy of the mental does not compromise its reality, for this insistence requires that the invariants – "what matters" – are mental. It appears that Davidson's position concerning indeterminacy is not stable.

This concludes my overview of Davidson's approach. The interpretivist elements in his work are an integral part of his larger meaning-theoretical enterprise. Therefore, it should come as no surprise that classifying his stance with respect to various interpretivist options is not a straightforward matter. Davidson did not explicitly aspire to work out an interpretivist account of mind; his emphasis on interpretability was rather motivated by his aim to provide appropriate truth conditions to a speaker's utterances. The measure of appropriateness was that these truth conditions should be part of the overall interpretation of the speaker. In order to find out what is sufficient for interpreting a speaker, and to avoid circularity, the interpretation should be radical, that is, it should proceed from an austere starting-point without presuming knowledge about the speaker's beliefs and truth conditions or meanings of his sentences. The emphasis of my project is different. Instead of trying to find out what the minimal knowledge-base for warranting ascription of beliefs and meanings might be, my aim is to formulate an interpretivist account of the possession of mental states, and this project need not involve setting radical restrictions on the availability of data. Nevertheless, Davidson's ideas concerning interpretation are a valuable resource and we return to them throughout this book.

3.3 Daniel Dennett and the intentional stance

In this section, I present and analyse another account that is commonly associated with interpretivism: Daniel Dennett's intentional stance approach. Dennett's

28. Cf. Rawling (2003: 104) who argues that for Davidson, the invariant nodes in different interpretation theories must be behavioural dispositions. On his reading, Davidson's position comes close to instrumentalism concerning the mental.

views have undergone several changes over the years. For example, he has presented different views on the status of folk psychology, ranging from instrumentalism to mild realism. His work has spanned many areas and topics, and it is not altogether obvious that his pronouncements on various matters have always been consistent. He also refuses to phrase his views in terms of categories and distinctions that are common among other philosophers. This is partly because he thinks that such categories fail to do justice to the multifarious nature of his subject and that working with clear-cut distinctions – "philosophy drawn and quartered" – may blind one to other, less distinct alternatives (Dennett 1993a: 216). Instead, he expresses his ideas in a characteristically vibrant style, using stories, analogies and intuition pumps rather than arguments consisting of numbered lines. Therefore, it should not be surprising that for many his position has seemed frustratingly hard to pin down.[29] In the following overview, I concentrate on what seem to me to be the basic elements of Dennett's account, focusing mainly on papers collected in *The Intentional Stance* (Dennett 1987).

A good point of entry into Dennett's account is his distinction between different ways one can explain and predict the behaviour of a system. Already in 1971, Dennett distinguished between different explanatory stances (Dennett 1971). He considers a chess-playing computer whose behaviour can be approached from the *design stance*: this is to predict computer's responses on the basis of its design. If one knows how the computer is supposed to function given its inputs, one can predict its outputs, given that the system performs its tasks as it is designed to do so. The design stance is not limited to the study of artificially designed systems. The behaviour of natural organisms can also be predicted from the design stance, given knowledge of how evolution has moulded their functions in their natural environments. Prediction from the design stance does not require knowing facts about the internal physical structure of the system, for presumably, to a certain extent, different structures can carry out similar functions.[30] However, physical structures become important in the next stance – the *physical stance* – that one can resort to when the system fails to perform its function and one seeks to explain its malfunctioning. From this stance, prediction is based on the physical state of the system and relevant laws of nature. The behaviour of the chess-playing

29. For some attempts to fit Dennett under one or another common label, see Jackson (1993), Fodor and Lepore (1993), McLaughlin and O'Leary-Hawthorne (1994).

30. One can distinguish between two notions of design (e.g., Elton 2003: 38). One notion is teleological and refers to the purpose of the system, while another notion understands design in an ordinary functionalist sense as a collection of a system's internal components and their interrelations. However, in Dennett's case, these two senses come together, for to understand how the program works, it is usually essential to know what the program is designed for.

computer can in principle be predicted from the physical stance, but in practice, the account would be unreasonably complex, given the ready availability of the prediction in functional terms. In addition, according to Dennett, when we predict the behaviour of the chess-playing computer from the physical stance, we fail to view this system as a functional unit that is designed to carry out a certain task. Rather, we see it as a physical unit following the laws of nature.

In some cases, when the design of the system becomes too complex, prediction from the design stance also becomes impractical. Then it is reasonable to take the assumption of rationality on board. For Dennett, this is to assume that the system selects the most optimal course of action. To take the rationality assumption is to take the *intentional stance*. Dennett (1971: 90) elaborates: "One predicts behavior in such a case by ascribing to the system *the possession of certain information* and by supposing it to be *directed by certain goals*, and then by working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions." The possession of information and having certain goals can also be characterised as having beliefs and desires. Predicting a system's behaviour from the intentional stance thus involves the attribution of folk-psychological notions of belief and desire to the system. Dennett's view is that any system whose behaviour can be reliably predicted from the intentional stance is an intentional system. In this sense, various artefacts like the chess-playing computer are intentional systems. This claim may seem like a stipulation on the technical notion of 'intentional system' and thus not as a particularly audacious statement. However, Dennett goes on to state that being an intentional system just is being a believer.

> I will argue that any object – or as I shall say, any *system* – whose behavior is well predicted by this strategy is in the fullest sense of the word a believer. *What it is* to be a true believer is to be an *intentional system*, a system whose behavior is reliably and voluminously predictable via the intentional strategy.
>
> (Dennett 1987: 15)

Elsewhere, Dennett identifies being a believer with exhibiting reliably interpretable behaviour:

> [*A*]*ll there is* to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence *all there is* to really and truly believing that *p* (for any proposition *p*) is being an intentional system for which *p* occurs as belief in the best (most predictive) interpretation. (Dennett 1987: 29)[31]

31. Dennett (1978: xvii) explicitly ties believing with the possibility of attribution: "(*x*)(*x* believes that snow is white ≡ *x* can be predictively attributed the belief that snow is white)."

This quote expresses several commitments of Dennett. First of all, there is an expression of the interpretivist element and the idea that belief is to be analysed from the third-person perspective, or from the perspective of the interpreter. Then there is the rejection of the distinction between original and as-if intentionality. A common view (e.g., Searle 1992: 78) is that even if we call the chess-playing computer an intentional system alongside ordinary subjects of folk psychology, there is a crucial difference between people and computers in that the former have original intentionality whereas the latter are intentional merely in a derived or metaphoric sense. One may concede that we can indeed say that the computer will make the following move, since the computer clearly desires to win and knows the distribution of pieces on the board, and this move is the most reasonable in the given state of the play, but this attribution should not be taken literally. However, Dennett would not agree. He rejects the very notion of original intentionality, since there is no explanation for its emergence from the non-intentional basis. The appearance of original intentionality should be resisted and the difference between people and artefacts should be explained without presuming mysterious notions that themselves are hard to explain. Dennett's point is thus that once we have explained how a chess-playing machine can have intentional states, there is no further task of explaining what it is for people to have intentional states.

Dennett is not exceedingly liberal with respect to ascribing intentionality either, for he allows that when the ascription of beliefs yields no predictive or explanatory results, the ascription is not justified. Hence, not everything goes: there are some standards to be respected. Lecterns do not have beliefs (Dennett 1987: 23). If the predictability of a system from the intentional stance fails to be "reliable and voluminous", it is not an intentional system. One may have certain qualms about whether this standard is precise enough, but in general, by using this standard, it is quite straightforward to confirm whether a prediction is successful or not. One simply has to wait for the predicted event to occur. In addition, as Dennett (1987: 29) himself is also quick to note, the stress on predictability actually places a constraint on the internal structure of a system as well. Although the attribution of belief does not proceed from facts about the internal constitution of a system, the display of large behavioural complexity would also presume a certain complexity of internal structure. It is likely that there are no direct specific entailments from belief attribution to internal organisation, and hence there are no true and strict psychofunctional laws. Nevertheless, the fact that a system's behaviour is predictable by beliefs and desires rules out certain kinds of internal organisations (like having no structure at all) and can give grounds for ascribing internal representations (cf. Dennett 1987: 33).

Another large issue raised by this quote from Dennett is a question about the reality of beliefs and other intentional states. A common worry is that Dennett is

a conservative eliminativist or an instrumentalist who holds that we do not really have beliefs and desires at all, but for pragmatic purposes it pays off to talk as if we had beliefs and desires. Indeed, in (Dennett 1971: 91), he says that taking the intentional stance is a pragmatic strategy. In "Three Kinds of Intentional Psychology", reprinted in (Dennett 1987), he characterises intentional states as *abstracta*, abstract constructs of a theory like centres of gravity, which are contrasted with *illata* that are concrete theoretical posits. However, later Dennett (1987: 72) came to regret calling his position "instrumentalism". This made it difficult for others to distinguish his position from other instrumentalist approaches, and hence it was assumed that he had inherited the common problems plaguing instrumentalism (Dennett 1993a: 210). Therefore, he preferred later to characterise his position as "mild realism". However, since the notion of realism has so many different connotations, it would make more sense to look more closely at which notion of realism is at issue here, and what characteristics intentional states have according Dennett, as well as which characteristics they lack.

Sometimes Dennett (1987: 71) opposes himself to Fodorian "Realism with a capital R", which assumes that the internal behaviour-causing states will turn out to have a taxonomy similar to that of folk psychology.[32] Elsewhere, Dennett (1988a: 496) has claimed that he is not quite the realist who holds that "beliefs are objective things in the head which could be discovered and whose identities could be confirmed, in principle, by physiological psychology". He does not think that it could be discovered that we have functional equivalents of beliefs and desires in the brain. His considered position is that by taking the intentional stance, an intentional pattern would become visible that is not accessible via other stances. Dennett's (1987: 25–26) example is that there are several ways one could buy stocks: one could place an order through the internet or raise a hand at the right moment in a right place and so on. When one describes merely the accompanying physical movements from the physical stance, one misses an objective pattern of buying stocks. Dennett (1991b) claims that these are "real patterns" that support predictions and generalisations – they are neither fictions nor reducible to physical patterns. He makes it clear that his position should rest on the middle ground between realism and relativism:

32. Fodor's (1985: 78) characterisation of Realism is a conjunction of the thesis of causal efficacy of the mental with the claim that these states have semantic properties: "I propose to say that someone is a Realist about propositional attitudes iff (a) he holds that there are mental states whose occurrences and interactions cause behaviour and do so, moreover, in ways that respect (at least to an approximation) the generalizations of common-sense belief/desire psychology; and (b) he holds that these same causally efficacious mental states are also semantically evaluable."

> [W]hile belief is a perfectly objective phenomenon (that apparently makes me a realist), it can be discerned only from the point of view of one who adopts a certain predictive strategy, and its existence can be confirmed only by an assessment of the success of that strategy (that apparently makes me an interpretationist). (Dennett 1987: 15)

Note that now the opposition is between the claim that propositional attitudes are objective phenomena and the relativist option (interpretationism in the sense of pure ascriptivism) that having mental states lies in the eye of the beholder. Given that 'realism' has been used in a variety of senses, the label itself fails to play a substantive role in the discussion. However, a certain characterisation of Dennett's position appears to emerge. Dennett is not a Realist in that he does not think that terms like 'beliefs' and 'desires' pick out causally efficacious and semantically evaluable entities that could be observed in the brain. However, this does not entail that beliefs and desires are not objective phenomena – they are not arbitrarily attributed. Neither does the account entail that someone "was a believer from the point of view of one observer, but not a believer at all from the point of view of another, cleverer observer" (Dennett 1987: 24). As Dennett puts it on the same page: "The decision to adopt the intentional stance is free, but the facts about the success or failure of the stance, were one to adopt it, are perfectly objective." The success-conditions of interpretation or whether the particular ascription is predictive do not depend on the interpreters. We will return to the question of objectivity later in this section when we discuss Dennett's views on indeterminacy.

Dennett has brought out several characteristics of propositional attitudes that presumably support their peculiar status. We have already mentioned his contention that beliefs and desires fail to line up with causal structures in the brain. In addition, beliefs have no clear-cut identity and possession conditions. Dennett (1988c) gives his list of reasons for assuming that beliefs are abstract patterns:

> There can be an infinity of beliefs in a finite brain; between any two beliefs there can be another; the question of exactly when one has acquired a particular belief can have no objective, determinate (interpretation-independent) answer (you weren't born believing that lawyers typically wear shoes, and you have believed it for some time – just when, exactly, did this belief get added to your belief store?) Dennett (1988c: 537)

Indeed, if (and this is a metaphysical point) beliefs have no fixed location in space and time and the subject cannot have a fixed number of beliefs, then they cannot be concrete entities.

He adds another reason: one can always learn something more about robustly real entities, whereas this is not the case for beliefs. The nature of mental states is fixed by folk psychology and once one masters folk psychology, no facts

about beliefs and their kin remain to be discovered: “what we can learn about them is just whatever follows from their combinational roles in our theories” (Dennett 1988c: 537). It might be objected that this fails to distinguish beliefs from other theoretical entities like electrons. It is true that once one knows the complete physical theory of electrons, there is nothing further to learn. However, there is a crucial difference. The nature of electrons coincides with our knowledge only at the end of the inquiry, and no additional explanatory value is gained by explicating the nature of electrons in terms of what an ideal inquirer would know about electrons at the end of his inquiry. After all, the realistic attitude (as opposed to constructivism) presumes that the nature of electrons would not change if we somehow fail to reach the end of inquiry or if we had not discovered electrons in the first place. The nature of propositional attitudes, however, if they are as Dennett envisages them, lies open to anyone who has mastered the common-sense theory of propositional attitudes, and an account of such attitudes would involve a story about their role in a common-sense theory as well as the mention of how they are employed in the everyday context. If such facts were omitted from the account of propositional attitudes, the account would essentially be incomplete.

A fourth kind of motivation that should be mentioned is Dennett’s claim that the intentional stance involves idealising assumptions about rationality. In his view, the attribution of beliefs and desires is guided by the assumption that the subject of attribution is a rational agent. The ascriber is supposed to attribute those beliefs the system “*ought to have*, given its perceptual capacities, its epistemic needs, and its biography” and the desires the system “*ought to have*, given its biological needs and the most practicable means of satisfying them.” It is further presumed that “a system’s behavior will consist of those acts that *it would be rational* for an agent with those beliefs and desires to perform” (all quotes from Dennett 1987: 49). The “ought to have” for a system is further unpacked in terms of being “ideally ensconced in its environmental niche”. The ascription-conditions of intentional states thus presume a strong form of rationality. However, there is a growing amount of empirical research showing that people actually fail to instantiate rational agenthood. Dennett has attempted to come to terms with this inconsistency between actual agents and idealised belief-ascription in two ways. Initially, in “Three Kinds of Intentional Psychology” (Dennett 1987: Chapter 3), this seemed to motivate an instrumentalist attitude towards folk psychology. If the latter is an abstract theory incorporating idealising assumptions, and mental states are nothing but the constructs of such a theory, then even if folk psychology is used to predict and explain our behaviour, it does not follow that we would actually possess beliefs and desires as they

are specified in folk psychology.[33] Ideal agents have beliefs and desires, actual agents merely approximate to the ideal. Since the difference between actual and idealised agents does not loom large for practical purposes, actual agents can be predicted from the intentional stance. But due to idealising assumptions, folk psychology is an instrument of explanation and prediction rather than a descriptive account (cf. Heal 2003: 231).

In later work, Dennett has stopped drawing instrumentalist conclusions from the rationality assumption. His later view (Dennett 1991b) is that even if the ascription of propositional attitudes is governed by rationality considerations, intentional systems are not required to instantiate ideal rationality. The suggestion now is that the intentional pattern may contain distortions and noise and that it can still be detected as a real pattern by an interpreter. However, the idealisation condition continues to play a role. It is in virtue of the noise approximating to the ideal or well-organised pattern that it becomes detectable from the intentional stance.[34] If the noise becomes prevalent, as may sometimes happen, then the pattern disappears. In other words, more reminiscent of Davidson, if the deviation from the norms governing beliefs and desires becomes too large, the interpretation would stumble. In addition, the rationality requirements are soothed (Dennett 1987: 94–98), although Dennett never explicitly says what the rationality is. This allows him to say that even noisy patterns (those that contain deviations from the ideal rational order) are still intentional patters, for then one need not be ideally rational in order to have propositional attitudes. However, the conditionality of the 'ought', the dependence on some rationality assumption would still be involved. (We return to the topic of rationality in the next section.)

Dennett allows the indeterminacy in the attitude ascription. Given the interpretivist identification of the ascription and the possession that Dennett (1978: xvii) endorses, this converts to the indeterminacy of the attitude possession. In Dennett's view, the indeterminacy is not just the Davidsonian idea of giving different descriptions to the same phenomenon, it is more radical prospect of there being "two different, but equally real patterns discernible in the noisy world" (Dennett 1991b: 49). He elaborates:

---

33. Dennett (1987: 52, my italics) writes: "It is the *myth* of our rational agenthood that structures and organizes our attributions of belief and desire to others and that regulates our own deliberations and investigations."

34. In Dennett's view, patterns are discerned in the process of abstraction and organisation of the data, and a pattern is never a one-to-one mapping of the available data (a so-called bit map): "a pattern exists in some data – is real – if there is a description of that data that is more efficient than the bit map, whether or not anyone can concoct it." (Dennett 1991b: 34.)

> [T]here could be two different systems of belief attribution to an individual that differed *substantially* in what they attributed – even if yielding substantially different predictions of individual's future behaviour – and yet where no deeper fact of the matter could establish that one was a description of the individual's *real* beliefs and the other not. (Dennett 1991b: 49)

Although in such a case, the choice between patterns would be directed by pragmatic reasons, prescribed by the context, this need not mean that relativism has entered into the attitude ascription. For it should be open to Dennett to maintain that both patterns belong to the subject and are to be found through the lenses of the intentional stance. Which pattern one chooses to highlight depends on the particular interests of the interpreter, but this does not have an influence on the matter of possession. To allude to a Dennett's (1987: 24) quote above: there is no need to say that one has certain beliefs from the point of view of one observer, but he does not have these beliefs from the point of view of another observer.

Dennett (1991b: 49) notes that if Davidson considered such a case and accepted that one can indeed be attributed substantially different beliefs, he "would have to grant that indeterminacy is not such a trivial matter after all." Davidson (1997: 120) rejects the invitation, since he does not regard predictability as a criterion for the possession of propositional attitudes, and he does not think that Dennett has provided evidence that "equally justified systems for attributing attitudes are incompatible". Davidson rejects the idea of incompatible systems, since there is no way to evaluate their incompatibility. It is true that Dennett has only entertained the possibility of substantial differences and he has not spelled out what can be the source of indeterminacy in cases, when there are different but equally acceptable interpretations. In presenting the ascription theory, I attempt to be more explicit in this regard.

Given Dennett's claim that one can discern objective intentional patterns from the intentional stance, the question about the metaphysical status of such patterns would immediately arise. Are they physical or intentional? If they are intentional, is Dennett committed to property-dualism or antireductionism? Lynne Rudder Baker (1989) has argued that Dennett's commitments present him a dilemma: if intentional patterns are not accessible from the physical stance, then Dennett should abandon physicalism, but if they are not something metaphysically distinct from physical entities, "the intentional stance seems, in principle, dispensable without cognitive loss" (Baker 1989: 313). In the first option, if patterns are nonphysical, then there is a pressing task of explaining their ontological status. In the second option, if Dennett wants to keep physicalism, then he cannot keep the claim that there exist real patterns that cannot be captured from the physical stance.

Indeed, Dennett has been reluctant to clarify his views on the ontology of patterns. He acknowledges that his views on his ontological commitments are "in happy disarray" (Dennett 2000: 356), but his general attitude is that ontological questions are either futile at the current stage, given that there is even no consensus on the ontological status of common objects like voices and bus-tickets, or that they are secondary in respect of our explanatory interests (Dennett 1993a: 212, 2000: 359). This attitude is in line with an interpretation of Dennett's approach as presenting a clarification of the role and nature of different explanatory practices (Viger 2000).[35] From the perspective that different stances are different explanatory practices with their own ontological commitments, the ontological status of beliefs and desires is no worse than the status of posits of physical theories, for they are all theoretical posits. The difference between them lies in their explanatory role – intentional states are only *abstracta*, whereas some posits of physical theories are also *illata*. In addition, this approach presumes that it makes no sense to aspire for a common ontology, unattached to a particular perspective, model or theory. However, the question concerning the interrelation of the posits of different explanatory stances would still remain open in this approach. Given that Dennett does not reduce the constructs of the intentional stance or the design stance to the constructs discerned from the physical stance, the ensuing pluralism or manifold of constructs would not be satisfactory from the ontological perspective, which attempts to picture the world as consisting of entities of few basic kinds. If the explanatory pluralism is substituted for ontology, then this may seem to outlaw certain questions that make sense in the standard framework. An example would be the question whether the reality is physical or not. Similarly, the notion of taking different explanatory stances toward the same system becomes problematic if the explanatory stance itself is regarded as securing the ontology of its constructs. There does not seem to be a stance-independent notion of the sameness of the system. If the intentional system is the same system as that approached from the physical stance, then the autonomy of the intentional stance would be compromised (cf. Elton 2003: 49).[36]

In my view, Dennett should confront metaphysical questions and I think that the ontological status of patterns requires clarification. Some of his claims support

---

35. Dennett (2000: 356) agrees with Viger's interpretation that he is not advancing an ontological thesis.

36. Elton's (2003: 50) suggestion is that the intentional stance is kept in the background even when we take the physical stance, for only from the intentional stance can one discern systems as intentional that can also be physically explained. However, this does not settle the ontological worries about intentional patterns.

the view that intentional states are to be analysed as behavioural dispositions, distinct from the categorical basis in the brain that realises them. Intentional states are ascribed mostly on behavioural grounds that are constrained by rationality assumptions. The patterns are thus discerned *in* the behaviour, and as dispositions to behave, they are supposed to provide an explanation for the observed behaviour or to help predict the ensuing behaviour (see McCulloch 1990; and McLaughlin and O'Leary-Hawthorne 1994 for this reading, supported by Dennett (1987: 25, 50) and Dennett (1991b: 46)). However, this does not yet answer the metaphysical question. What are the behavioural dispositions that do not reduce to their realisers? If this is Dennett's view, then it is subject to objections similar to those directed at functionalism in the previous chapter. But I think that a more plausible interpretation of Dennett is to note that he understands dispositions to behave in a very extended sense: they include everything that can be manifested by the system. They are not just dispositions to peripheral behaviour but also "everything that happens in the brain" (Dennett 1993a: 211). On this reading, the dispositional patterns are just any patterns of the system that are apt to appear or become manifest to interpreters. However, this reading does not settle their metaphysical status either.

I propose that the approach to metaphysics outlined in the previous chapter might prove useful in this regard. We saw that the shape of the worry concerning the relation between beliefs and physical entities remains even if one deflates ontology into explanatory practice. It may, however, be helpful to keep matters of explanation and ontology apart. This is to draw on the distinction between pleonastic entities and elements of the natural basis. The former play a role in our various explanatory projects, and the latter make up the ontology. Now we can say that mental entities, being merely pleonastic, do not have the same ontological status as the elements of the natural basis, whatever those might be. This distinction also allows us to give an answer to Baker's dilemma. She set up the issue as a choice between intentional patterns being nonphysical (if not reducible to the physical) or being dispensable (if reducible to the physical). This dilemma conceals the distinction between the conceptual and the ontological reduction. If we rely on this distinction, which to a certain extent mirrors the distinction between the pleonastic and the natural, then we can say the following. Intentional patterns are pleonastic; they form a distinctive conceptual and explanatory order. They are not (conceptually) reducible to physical pleonastic properties, so the intentional stance is not dispensable. This does not entail that the intentional patterns are nonphysical from the ontological point of view. From that perspective, they do not enter into the view at all, since there are no inflationary mental properties. The mental (or the intentional) is not an ontological category.

Still, one may suspect that I am opting for the second horn of Baker's dilemma: that intentional patterns are "dispensable without cognitive loss, in which case

Dennett's intentional stance is 'just a sham and a word game' (Haugeland 1982: 616)" (Baker 1989: 313–314). This is not quite right. Indeed, intentional patterns are dispensable without *ontological* loss, but the ensuing cognitive deficit would be no less than the loss of folk psychology and the adjacent concepts by which we understand ourselves and our kin. Moreover, they are real patterns, if real is what is objective. Being pleonastic does not preclude intentional patterns from being objective in the sense that any actual observer can miss them. This is one suggestion for an answer to Baker's dilemma, and in principle, this sort of answer would be available to Dennett, if he were prepared to share its underlying metaphysical assumptions.

There is another important line of critique directed against Dennett's account (see McCulloch 1990; Lyons 1995: 28–29). How does Dennett explain the truth and success of behavioural predictions? He cannot say that these predictions are successful because they involve reference to those beliefs and desires that have played a role in the causation of behaviour, since that would presume that beliefs and desires have actually caused the predicted behaviour. In his view, folk psychology does not point towards the internal states that cause behaviour. But then, why do predictions based on folk psychology work so well?

To answer such questions, Dennett introduces evolutionary considerations. Basically, his idea is that folk psychology, as well as the discovery of the intentional stance, is what has been selected by natural selection during long-term trial-and-error testing. The numerous false starts that were not selected have disappeared from the picture and the most predictive method of interpretation, which incorporates rationality and optimality constraints, has remained. Hence, folk psychology does not work because it gives an accurate picture of what is going on in our brains, but rather because it has proved to be adaptive in the long run: it "has evolved because it works and works because we have evolved" (Dennett 1987: 48–49). The second part of the last claim points to Dennett's view that "evolution has designed human beings to be rational, to believe what they ought to believe and want what they ought to want" (Dennett 1987: 33).[37] In this respect, folk psychology with its rationality assumption and rational folk are tailor-made to each other.

---

37. Note that Dennett (1987: 314ff) adopts the intentional stance towards evolutionary functions as well. Some (e.g., Baker 1989; Heitner 2000) have thought that this conflicts with the role that evolutionary considerations play in his account, but there is no inconsistency here. The brute and blind processes of natural selection are particular and non-intentional causal processes. Thus Dennett (1990b: 190) says that the evolutionary account "shows how every feature of the natural world can be the product of a blind, unforesightful, non-teleological, ultimately mechanical process of differential reproduction over long periods of time." However, when we describe the process of natural selection from the intentional stance, we can begin to make sense of it in terms of purpose, optimality and biological function.

Let us now focus on the question of whether interpretability or the interpreter is ineliminable from Dennett's account. McLaughlin and O'Leary-Hawthorne (1994) have analysed Dennett's views by looking for the supervenience basis for being a believer. They proceed from Dennett's claims that connect believing with behavioural dispositions, as in the following statement:

> I am claiming, then, that folk psychology can best be viewed as a sort of logical behaviorism: *what it means* to say that someone believes that *p*, is that that person is disposed to behave in certain ways under certain conditions. (Dennett 1987: 50)

On such grounds, McLaughlin and O'Leary-Hawthorne (1994: 208) argue that "being a genuine believer supervenes, on Dennett's view, on dispositions to peripheral behavior. Stance-takers are not part of the supervenience base." However, neither they nor Dennett distinguish between explaining what belief or believing is (what 'belief' or 'believing' mean) and explaining what it is to possess a belief.[38] If we make this distinction, then there is room to argue that the nature of belief supervenes on behavioural dispositions, whereas stance-takers are part of the base on which possession supervenes. Proceeding from Dennett's claim in the quote above, one can say that stance-takers enter into the explanation of what it is to possess dispositions to behave.

William Child (1994: 50–51) has also presented reasons for thinking that Dennett's view is a version of non-constitutive interpretivism. If what is involved in having beliefs and desires in his account is spelled out, he argues that references to an interpreter or the intentional stance can be deleted. Child adds together several claims of Dennett, (like those in 1987: 15 and 49, both quoted above), and derives the following statement:

> *What it is* to be a true believer is to be an intentional system, a system which behaves in a way that would be rational for an agent which had the beliefs and desires that system ought to have in the light of its circumstances and needs. (Child 1994: 51)

This elaboration of what it is to be a believer does not refer directly to interpretability. However, as in the previous case, one can argue that interpretability is implicitly contained in the statement, for the facts concerning the beliefs and desires an agent ought to have in his situation are determined by facts involving interpretation.

38. Note that although the meaning of 'X has belief that *p*' differs subtly from the meaning of 'to believe', it does not yield an account of what is it to possess the belief, which is an altogether different matter.

Indeed, Dennett has not been too explicit in this regard, but I would argue that he needs to make some such claim if he is wants to distinguish his position from more common brands of behaviourism or instrumentalism. Otherwise, his view would amount to something like this: one's having mental states is a matter of the possibility of explaining and describing one's behavioural dispositions in intentional terms. In this picture, the intentional stance becomes an epistemic device which detects intentional patterns that are otherwise simply difficult to discern. The interpretivist element would thus be reduced to mere revelationism. Indeed, Dennett has expressed something suspiciously close to this view:

> [W]hat makes it true that people have the *real* mental states is facts about their *behavioral* dispositions and capacities, but these facts can be perspicuously and efficiently expressed only from the intentional stance, an *instrument* of prediction (and explanation). (Dennett 1993: 923)

On a superficial reading, this would again give rise to worries about the status of intentional patterns. For if what is really there is behavioural dispositions, then in what sense can one say that intentional patterns are mildly real? Instrumentalism concerning intentional facts seems to threaten again.

However, there is a different reading. It points towards a position close to the ascription theory, and I put it in terms I would countenance. First, using the pleonastic notion of a fact, we can distinguish in the quote above between facts about behavioural dispositions and intentional facts; both facts are pleonastic, being matters of a true description. But this is not to embrace selective instrumentalism concerning the mental, unless we avail ourselves of an inflationary notion of a fact, which we should not. Now consider the claim that an intentional pattern can only be expressed from the intentional stance. This is to say that intentional descriptions can only be given in intentional vocabulary. And this is the thesis of conceptual irreducibility: there is no way to characterise what is described in intentional terms in some other vocabulary. Further, let us add the claim that the application of intentional vocabulary requires interpretation. It can be argued that the interpreter cannot be deleted from the account of what it is to assign mental descriptions. The application conditions of mental terms must include reference to an interpreter, since the applicability of a given mental term is dependent on whether one can be recognised as meriting a description in these terms. To put it in Dennett's terms, mental patterns are dependent on their recognition as mental patterns.

There is some textual evidence that Dennett does actually hold such a view (cf. Seager 2000: 117). For example, in explaining the sense in which mental patterns are objective, he says:

> These patterns are objective – they are *there* to be detected – but from our point of view they are not *out there* independent of us, since they are patterns composed partly of our own "subjective" reactions to what is out there.
> (Dennett 1987: 39)

In "Real patterns", he claims that patterns are closely connected to their recognition:

> [P]attern is "by definition" a candidate for pattern *recognition*. (It is this loose but unbreakable link to observers or perspectives, of course, that makes "pattern" an attractive term to someone perched between instrumentalism and industrial-strength realism.) (Dennett 1991b: 32)

Formulated in terms of mental patterns, the idea is that the presence of intentional patterns is objective in the sense that the correctness of an ascription does not depend on the interpreter, but the presence of patterns is still recognition-dependent, since patterns are "recognizabilia", to use Haugeland's (1993) memorable term.[39] In conclusion, contrary to Child and to McLaughlin and O'Leary-Hawthorne, there is therefore a reading of Dennett as a constitutive interpretivist, such that interpretability is ineliminable from his theory. We can say that on this reading, his position conforms to intermediate interpretivism, as specified in Section 3.1.

The recognition-dependence of intentional patterns can be used to counter a critique that is often directed at interpretivist approaches (e.g., McLaughlin and O'Leary-Hawthorne 1994: 227; Lance and O'Leary-Hawthorne 1997: 274). This was mentioned at the beginning of this chapter. Specifically, it has occurred to many that the interpretivist project of explaining the possession of mental states and contents in terms of their ascription is guilty of conflating metaphysics with epistemology. In place of a metaphysical account of beliefs or what facts are involved in having a belief, we are offered an account of how one could come to know what the subject believes. It seems that no distinction is made between the evidence we have that someone believes that *p* and what constitutes someone's believing that *p*. In some cases, such a move is clearly confused: compare this with explaining how we come to recognise pieces of gold when the explanation of the chemical structure of gold is required. However, there are considerations that explain why metaphysical and epistemological considerations cannot be separated in the case of mental states and their contents. If mental properties do not have a deep metaphysics in the sense that there are no inflationary

---

39. Haugeland (1993) finds Dennett's different definitions of 'pattern' to be a source of conflict. Dennett (1993a) denies that there is any conflict between his claims that patterns are objective and yet recognition-dependent.

mental properties and there are no facts about the mental that are constituted by something requiring discovery, then epistemological and conceptual considerations take the place of the metaphysics of mind. All that has to be explicated are the meaning and application conditions of mental terms. Hence, if (pleonastic) mental properties are recognition-dependent in the sense explained above, then it should be acceptable to refer to recognition and other epistemic matters in a constitutive account of such properties. Note that this line of response is not shared by Davidson, since he does not think that mental events are recognition-dependent. He holds onto physicalist metaphysics with respect to mental tokens, and gives an epistemic account of mental types. Apparently, he assumes that there is nothing metaphysical to say about mental types. However, as pointed out earlier in this chapter, his account is in principle compatible with constitutive interpretivism about mental types.

In sum, of the different readings that one could give of Dennett's account, I have presented and supported my own interpretation of his views and made some suggestions as to how, given this interpretation, one may answer various objections. To my mind, his major insight is that although intentional states can be discerned only from a certain perspective, and are thus recognition-dependent, what is discerned does not thereby fail to be objective. This basic idea is shared by the ascription theory as well.

While Davidson's approach was driven by meaning-theoretical considerations, Dennett's interpretivism is motivated by the contemplation of folk psychology. He strives to make sense of mentality on the condition that there are no internal causally efficacious mental states mirroring the taxonomy of folk psychology. There are several further dissimilarities between Dennett and Davidson. Their different positions on the indeterminacy and the reality of the mental were already discussed, but there are more divergences. It is beyond this book to discuss all the differences here, but I mention briefly some of them. Dennett certainly does not follow Davidson in espousing token identity. Davidson in turn does not say that beliefs and desires are idealisations and he does not rely on evolutionary considerations. He is less liberal than Dennett in ascribing intentional states. While for Dennett, taking the intentional stance is justified once it proves to be predictive, Davidson seems to lay stricter criteria on the possession of intentional states. Mainly in his later work, collected in Davidson (2001c), he has explicated several transcendental or necessary connections between thought and the concept of belief, the ability to be surprised, communication, language and the ability to have the concept of an error or the ability to make sense of the objective-subjective contrast. I do not discuss these more ambitious arguments in this book, for they go too far from what the ascription theory is prepared to claim.

3.4 Some roots of divergence

In the final section of this chapter, I discuss some aspects on which my particular interpretivist approach differs from those of Davidson and Dennett. I also discuss some objections to their views that I do not treat elsewhere. There are various objections that I do not touch on here: e.g., the circularity charge and various thought experiments involving Blockheads, Martians and so on. These will be countered in Chapter 6, from the standpoint of the ascription theory. Here I consider three main issues: the problem of rationality, the evidence used in interpretation and the role of the explanation of meaning.

3.4.1 Rationality

A common and recurrent argument against Dennett and Davidson is the following. Interpretation must presume that people are mostly rational. But people are not mostly rational. Hence, interpretivism gives us a skewed picture of human agents. It cannot account for the irrationality and mistakes of people. More sophisticated versions of this basic objection are given by Stich (1981, 1990), Thagard and Nisbett (1983), and Goldman (1986) among others.

As we saw before, Davidson counters this objection by saying that the attribution of irrationality is feasible only on the assumption that the subject is to a large extent rational, for only in light of rationality can we make sense of the occurrences of irrationality:

> Rationality is a matter of degree; but insofar as people think, reason, and act at all, there must be enough rationality in the complete pattern for us to judge particular beliefs as foolish or false, or particular acts as confused or misguided. (Davidson 1984a: 18)

To this, it has been objected that people can have special delusional thoughts that cannot be rationalised in light of their other beliefs and desires. These are states that one can fall into without a sufficient reason and one could stick to them also in cases when there is good evidence to the contrary (Bortolotti 2005: 202).

Dennett's pronouncements on the topic of rationality have not been satisfactory either. He argues that intentional systems are largely rational, since only rational systems have been selected during the course of evolution. This should also explain why our folk psychology is driven by the rationality assumptions. Dennett (1987: 84) has also suggested that irrational behaviour cannot be predicted from the intentional stance or the prediction would be indeterminate. To predict irrationality one has to view the subject as a malfunctioning system and this means that one has to descend into the design stance. However, this is contrary to the

common practice of ascribing irrational beliefs in folk psychology. Dennett's evolutionary considerations are also open to critique, for there is no reason to assume that only rational beliefs have selective advantage and serve the purpose of survival (cf. Stich 1981; Child 1994: 51, n. 93). In reply to his critics, Dennett has attempted to clarify his notion of rationality. He distinguishes it from stronger notions of deductive closure and logical consistency and notes that, on pain of a vicious circle, he cannot "define rationality in terms of what evolution has given us" (Dennett 1987: 96). He proposes a weaker notion of rationality, which is merely contingently related to "the methods of getting ahead, cognitively, in the world" (Dennett 1987: 97). However, this is too unspecific to back up his other claims on the relation between rationality and intentionality.

I am inclined to think that the quarrel over rationality is a red herring in discussions about interpretation. As the interpretivist approach is also possible without the assumption that people are rational, we can make progress by shifting the focus from the rationality issue to what really matters for interpretivism. What really matters is whether a subject's mental states and contents can be inferred or constructed on the basis of the available evidence. The possibility of this presumes that mental states can be somehow related to the sources of evidence accessible to the interpreter. For the ascription to be achievable, there must thus be a way of positioning the subject in relation to the ascription sources so that some mental description becomes applicable. One can formulate some principles pertaining to each ascription source, but these are all defeasible – each of them can be overridden by some further relevant facts. For example, although there is a presumption that people have true beliefs about their surroundings, the ascription would not proceed on the basis of this presumption if there were compelling evidence for ascribing illusory or hallucinatory contents to the subject. As outlined in Chapter 5, one of the additional requirements set by the ascription theory is that the ascribed state is in harmony with such compelling evidence. But it is not demanded that only such beliefs and desires that make the subject internally consistent or coherent are attributable to the subject.

Janet Levin (1988) has discussed various cases of irrational action that can all be characterised in mental vocabulary.[40] The ascription theory is capable of accommodating such situations. One such case is the wishful thinker, who believes that not-*p*, although he is in the perceptual contact with *p*. The belief that not-*p* can be ascribed to the thinker in conjunction with the dislike of *q*, and the belief that *p* produces *q*. In that case, the interpretation would follow the principle: "If *x*

40. More recently, Bortolotti (2004) has also argued that in ordinary cases, we can interpret irrational behaviour in intentional terms. Namely, we can do this with the help of background knowledge and heuristics, without needing to rely on the overarching rationality assumption.

observes that *P* and believes that *P* brings about *Q*, and desires that *Q* not occur, then *x* will believe – *P*." (Levin 1988: 202.) This principle is not rational, but it uncovers a way of making the agent intelligible by applying mental terms to him. Among her other examples is a creature acting against the principle of induction and a person who constantly comes to false perceptual beliefs. As concerns practical rationality, she refers to people who act against their best interests and to the possible agent who tries to avoid getting what he most strongly desires. The latter agent violates the practical syllogism that if one desires *q* and believes that doing *p* would result in *q*, and can do *p*, one does *p*. Since we can make sense of his behaviour in light of the fact that he does not follow his strongest desires, the agent's failure to exhibit practical rationality does not preclude the possibility of interpretation.

In another colourful example of irrational behaviour, Levin imagines someone who wishes to clean a carpet and believes that by using a vacuum cleaner, one can achieve the best results. However, instead of cleaning the carpet, he washes his car. This is puzzling, even more so as the carpet badly needed cleaning, whereas the car was freshly washed. Everything falls into place once it becomes evident that thinking about vacuum cleaning reminds this person of car washing and that in general, his behaviour accords to an idiosyncratic regularity: "If it wants *Q*, and believes that *P* brings about *Q*, and *P* reminds it of *R*, then it will do *R*" (Levin 1988: 212). Again, we can get a handle on this creature by interpreting its behaviour by taking it to conform to this idiosyncratic regularity or principle. Note that this principle relates beliefs and other intentional states to the creature's behaviour. Hence, this case can still be described by using mental vocabulary and mental states can be ascribed to the creature.

The picture of interpretation that emerges from Levin's considerations is one that aims to make its subject intelligible. As long as one's behaviour can be described in mental terms in view of the available evidence, mental states are ascribable. This presumes a distinction between the interpretation that makes the subject intelligible and the interpretation that makes the subject rational. The latter notion is much stronger, and given the widespread phenomenon of irrational but mentally describable creatures, it cannot be a constitutive requirement for interpretation.

Thus far I have talked about what is minimally required for interpreting others in mental terms. One may wonder if there are also any assumptions that should be made about the subject of ascription. Levin (1988: 206) formulates some very weak conditions that the creature or the system should meet in order to be ascribed mental states. The beliefs of the creature should "arise systematically from perception", its desires should "arise systematically from its goals and needs", the beliefs and desires should "produce behaviour according to (sufficiently) stable

and lawlike principles", and the principles should point to the relations between the contents of beliefs, desires, perceived situation and behaviour. These assumptions are compatible with the ascription theory. I would only add that they should be read as assumptions imposed on the intentional system, not as a claim that the system has beliefs and desires that are fundamentally ascription-independent. If this is what needs to be assumed about the subject of ascription, then crucially, it is not presumed that the subject is rational in any strong sense. Levin rejects this stronger sense thus:

> According to these conditions, however, perceptual beliefs need not be *reliable*, inductive procedures need not be even minimally *truth preserving*, and the principles that describe the influence of belief and desire upon action need not approximate the canons of decision theory. Nor must these principles be adaptive, or evolutionary sound. (Levin 1988: 206)

The conditions listed above are sufficient to invite the ascription, but they are not necessary, for even if in some extraordinary case we cannot assume that there are systematic links between a subject's beliefs, environment and behaviour, there could still be facts explaining the lack of those connections. Given such facts about these non-systematic links, some mental profile would then still be ascribable.

The approach I am recommending retains a version of the Dennettian sense of *ought* in the ascription (one ought to have such and such mental states, given the evidence!), and modifies the Davidsonian assumption concerning the consistency of a subject's beliefs. Instead of imputing internal consistency of thought and rationality to the subject, the ascription proceeds from the external sources of evidence about the subject. As will be explained in Chapter 5, the evidence from these sources should cohere, that is, it should point in the same direction, but this coherence can be compatible with ascribing internally incoherent thoughts to the subject. Indeed, such an account is incompatible with the view of radical interpretation that conceives it as proceeding from the scant basis of charity and rationality assumptions. It may indeed be the case that given the limited amount of data available to a radical interpreter, interpretation has to begin with some rationality assumption (cf. Levin 1988: 205). However, this is not an issue for the ascription theory, which does not assume that radical interpretability is constitutive of the possession of mind.

3.4.2 The sources of evidence for interpretation

The previous discussion of rationality leads naturally to the question about what sources of evidence there are for interpretation. It is an important part of my

account that different sources of evidence[41] must fit with the interpretation of the subject. If there is only one main evidential source, namely behaviour, talk about different sources being in harmony becomes empty. Therefore, my approach involves expanding the sources of evidence that are relevant in interpretation.

The radical interpreter proceeds from the subject's behaviour, his interactions with the environment and the sentences the subject holds true. I have already criticised this approach for requiring but failing to start from the non-semantic base. The point is that once we allow semantic facts as evidence for interpretation, there is nothing to stop us allowing additional facts about the person to be interpreted. It is often argued that the usual behavioural facts underdetermine the interpretation of a subject. Alex Byrne (1998) has described several situations in which this is the case.

> [S]uppose that someone is looking in good light at a cup on a table, affirms that there is a cup on the table, reaches out for it when the tea is ready to pour, and so on. Clearly he believes that there is a cup on the table! But not necessarily. He may take himself to be the victim of some crafty cup-illusion, and not wish to give the impression that he is wise to the trick. (Byrne 1998: 219)

He argues that one must be able to rule out such hypotheses, but this requires that one must "know what the subject *would* have done in various circumstances, not merely what he *actually* does" (Byrne 1998: 219). Another case he describes involves a person glimpsing a rabbit hidden in grass and having the thought that there is a rabbit, but behaving as if one had not noticed the rabbit at all. Given the facts about his behaviour, one can as well assume that the person did not notice the rabbit at all. Byrne argues that this thought could have influenced one's behavioural dispositions, for example, he could have reported that there was a rabbit, although he never did. Byrne (1998: 213) calls this the *problem of actuality* – "knowledge of how the subject *actually* behaves – both inside and outside the head – will not deliver the right results." To solve this problem, the interpreter should have "knowledge of counterfactuals that *cannot* be known on the basis of how the subject actually behaves." However, the grave problem with this "solution" is that ordinary people cannot have knowledge of such counterfactuals.

I agree that the ascription cannot be based on inaccessible counterfactuals concerning the subject's behaviour. Indeed, such cases present a good objection

41. The term 'evidence' might lead one to the thought that it is *evidence for* something that exists independently from the evidence itself, and this goes against the recognition-dependence of mental states. The idea is rather that it is evidence about behaviour, evidence about environment, etc., that forms the basis of ascription. Therefore, I would prefer to talk about the sources of ascription.

to those interpretivist accounts that attempt to proceed from the slender basis of actually exhibited behaviour. There are several points that should be made here. I think that the cup-illusion case can be captured if one concedes that idiosyncratic facts about the believer matter in interpretation. They determine which alternatives are relevant enough in order to be ruled out. However, in an ordinary situation, such options need not be ruled out, since there is a presumption that people are not normally subject to illusions. Another of Byrne's examples concerns hidden thoughts that are not behaviourally manifested. This case should be treated separately. First, there is always the prospect of self-ascription. We can still say that one has the thought, if the thought can be self-ascribed. Of course, this presumes an interpretivist account of self-ascription. One such possible account is presented in Chapter 9. Second, if the thought cannot even be manifested to self-ascription, then the coherence of the case becomes dubious. Such events are better described as protocontent and not as thoughts (see again Chapters 8 and 9). Another thing to note is that mental states are ascribable in holistic bundles. In order to ascribe a certain belief, some other states should also be ascribed in order to make the belief intelligible. So if the fact that one saw the rabbit is a part of the explanation of some later behaviour, then it is ascribable *post factum* as a part of a larger mental profile.

The general lesson from Byrne's examples is that interpretation is underdetermined by a subject's behaviour (including sentences held true) and his environment. This can be fixed by allowing more sources for the ascription. Note that this option is open to Dennett but not to Davidson. Dennett (1987: 49) indeed includes the "perceptual capacities" and "biography" of a system among the conditions that ought to be considered when ascribing beliefs.[42] Davidson may have wished to take the individual peculiarities of subjects into account at a late stage of the refinement of the theory of interpretation for a speaker (see e.g., Davidson 1974: 320). However, Lepore and Ludwig (2005: 218–220) point out that Davidson cannot draw on empirical and contingent facts about believers' idiosyncrasies, since he aims to deduce general considerations from his account of radical interpretation. Otherwise, his conclusions would only apply to some restricted class of speakers rather than to meaning and thought in general.[43]

---

42. Other philosophers have also proposed that particularities about persons play a role in their interpretation. Thus David Lewis (1974) brings in the subject's whole physical "life history of evidence" and Føllesdal (1982: 316) suggests relying also on facts about "past experience" and "personality traits, such as credulity, alertness, reflectiveness" of the subject.

43. Lepore and Ludwig (2005: 338) note that Davidson's reliance on history in the case of his Swampman's thought experiment also conflicts with taking radical interpretation as foundational in ascribing content.

3.4.3 Explaining meaning

Perhaps the most crucial difference between Davidson's approach and the ascription theory is that the ascription theory is not presented as a theory of meaning. In Davidson's approach, both a subject's beliefs and the meanings of his utterances are fixed in the process of (radical) interpretation. It is assumed that the subject's beliefs conform to the principle of charity and this allows for the assignment of interpretive truth conditions to the subject's sentences. For several reasons, I do not think that this can provide a complete account of meaning. Of course, it would be a separate (and massive) task to provide a proper critique of Davidson's theory of meaning. What I can do here is list some reasons why the ascription theory diverges from Davidson's approach and bring out the differences between these accounts. Some hints towards alternative ways of analysing meaning are given in Chapter 5.

One reason for the divergence is that I have assumed a deflationist position concerning truth. It would be uninformative to analyse meaning in light of a deflationist account of truth. This has been pointed out already by Dummett (1959). If we explain the 'is true' in a T-theorem:

> 'Snow is white' is true$_L$ if and only if snow is white

in terms of some deflationary understanding of truth (e.g., as in Horwich 1998b):

> The proposition that *Snow is white* is true if and only if snow is white

then we have not provided an illuminating account of truth that would be substantial enough to explain the meaning of 'Snow is white'. As a result, both the notion of meaning and the notion of truth would turn out to be insubstantial notions. The deflationary account of truth should therefore be coupled with a substantial account of meaning. Of course, deflationism about truth is not an obligatory component of the ascription theory, which can be presented also in terms that give a central role to substantial notions of truth and truth conditions. Hence, this argument should be viewed rather as a reason for searching for an alternative explanation of meaning, conditional on the deflationist premise.

Another reason for the divergence is that Davidson apparently downplays the use or importance of the notion of expression-meaning. He takes meaning primarily in the sense of speaker-meaning: it is what the speaker means by uttering something. Taken in this way, meaning something by some utterance is like a kind of mental state: it is a complex intention to be understood in a certain way. Davidson is explicit on this point:

> What matters to successful linguistic communication is the intention of the speaker to be interpreted in a certain way, on the one hand, and the actual interpretation of speaker's words along the intended lines through the interpreter's recognition of the speaker's intentions, on the other. (Davidson 1990b: 311)

In this way, meaning is made dependent on intentions.[44] This is reminiscent of Grice (1957), whose influence Davidson acknowledges, but it is also importantly different from the Gricean approach. Namely, Davidson is far from attempting to *reduce* meaning to nonlinguistic mental contents of intentions. As already explicated in Section 3.2, he explains how both meanings and beliefs are fixed in the process of interpretation. His contention is that neither meaning nor mental content can have explanatory priority with respect to each other: "the interdependence of the basic intentional attitudes is so complete that it is bootless to hope to understand one independently of understanding the others." (Davidson 1990b: 316.) His emphasis on a speaker's intentions does not conflict with the approach to meanings through truth conditions either. Truth conditions presume a speaker's intentions: "An utterance has certain truth conditions only if the speaker intends it to be interpreted as having those truth conditions." (Davidson 1990b: 310.)

This sort of speaker-meaning can indeed be specified along with a speaker's other beliefs, and there is no reason why the ascription theory cannot accommodate speaker-meaning. A speaker's intentions are ascribable in similar conditions as other mental states. But it is important to notice that speaker-meaning cannot be all there is to the notion of meaning. The notion of meaning has theoretical roles that cannot be served solely by speaker-meaning understood as above.[45] The role that is central in the present context is related to the question of the circularity of the interpretivist account. Namely, in order to ascribe mental states one has to presume that there are mental state terms with more or less fixed meanings, otherwise one would lack a language in which to carry out the interpretation. It is hard to see how such meanings could come from the intentions of the subject that is being interpreted. It is one thing to fix the content of a speaker's beliefs in view of what he intends to communicate, but it is another thing to fix the meanings of such terms as 'believes', 'intends' and so on. These meanings or the intentional vocabulary have to be assumed before interpretation begins.

44. Elsewhere, Davidson (1994b: 12) claims that "the concept of 'the meaning' of a word or sentence gives way to the concepts of how a speaker intends his words to be understood, and of how a hearer understands them."

45. For one potential role see Kemmerling (1993), who argues that the condition of being a normal speaker of a language is a part of the general criterion of belief-ascription, and therefore the notion of a shared language (which cannot be based on speaker-meaning) has philosophical import.

Another way to put the same problem is the following. Davidson's procedure gives us a list of T-theorems that presumably specify the meanings of a subject's sentences, or, in other words, suffice to yield the understanding of a speaker's utterances. But this creates the need to explain the meaning of T-theorems themselves or what it is to understand them. Horwich (1998a: 72) points out that "even if one accepts the Davidsonian picture, it remains to be said what it is for a sentence to have a certain truth condition: how does this come about?" Thus, an interpreter has to presuppose the meaningfulness of his own language used to formulate the T-theorems.

One can agree with this but point out that it is merely to assume that the interpreter has a language and that this language should in turn be radically interpretable. But that would lead to an infinite regress in which the interpretation of a language would in turn require interpretation. I do not see how Davidson could avoid such regress, and perhaps he would agree that there is no privileged or prior standpoint for explaining meanings and intentional states, for that would constitute a possibility for reduction. In some sense, the ascription theory is also faced with a circularity charge – namely, the ascription itself is a kind of mental state, and hence no reduction of intentionality is effected. I discuss this objection fully in Chapter 6, but here it should be noted that these types of circularity are different. Given the assumption that the intentional vocabulary cannot be reduced to non-intentional vocabulary, a certain regress is to be expected, if one explains intentional states in intentional terms; this does not show that the account cannot be illuminating, provided that there are separate stories to tell about the meaning of mental terms and the metaphysics of mind. However, the regress that faces the Davidsonian approach is more severe, for Davidson hopes to accomplish more – to explain both belief and meaning using the same apparatus. If it is objected that his account would in turn presume meaningful terms, it is not enough to point to a further interpreter that would interpret the utterances of the first interpreter.

This is one reason why I have not presented the ascription theory as an account of expression-meaning. This account is to be provided independently from the ascription theory. In Chapter 5, I will discuss further which potential forms of such account are consistent with the ascription theory, but presenting a full account of expression-meaning is beyond the present book. For now, the basic contention is that the account of expression-meaning can be prior to or independent from the ascription theory, given that meaning is explained in terms that do not presume mental states that are constitutively dependent on ascription. It is clear that the account of speaker-meaning cannot be prior to the ascription theory, since speaker-meaning is explained in terms of intentions, which have to be ascribable. Since the ascription of speaker-meaning thus presumes a background of

more or less fixed expression-meanings, there is no way the expression-meaning can be explained in terms of speaker-meaning in the present framework.

My distinction between explaining the nature of belief and explaining the possession of belief also reflects a desire to mitigate the circularity charge. If the ascription theory explains possession, it presumes the concept of belief which is part of the common intentional vocabulary brought together by folk psychology. In ascribing beliefs, we thus reach out to the common meanings of mental terms, which are explained separately.

In this chapter, I introduced and discussed two main interpretivist accounts – Davidson's radical interpretation approach and Dennett's theory of intentional systems. Their approaches have influenced me to a great extent, and this chapter provides necessary background for relating my own account to these established forms of interpretivism. However, I also brought to light several main principled roots of my divergence from their approaches. These are connected with the centrality of rationality in interpretation, (mainly in case of Davidson) the evidence there is for interpretation and whether there is a notion of meaning presumed by interpretation. Besides these departures, our accounts of mental causation, self-knowledge and perception differ as well. Some of those differences are discussed in the following chapters.

PART II

Elaborating and defending the ascription theory

CHAPTER 4

Folk psychology and mental terms

In the remaining parts of this book, I focus on presenting my own positive views. The chapters in part two form a unity: the present chapter discusses the requirements of the ascription theory from folk psychology, the next chapter outlines the basic position of the theory itself and Chapter 6 counters various potential objections. The chapters in part three discuss mental causation, perception and self-knowledge from the interpretivist point of view.

Part of the argument for interpretivism depends on certain views about folk psychology, namely, that concepts of mental states are constituted by folk psychology. In this chapter, I spell out how this locution should be understood. In the first section of this chapter, I outline the popular view that folk roles provide meanings of mental terms. This is the thesis of common-sense functionalism. I distinguish my approach from functional definitions and provide additional criticism of functionalism. The overarching idea of the present book is that having mental states is a matter of ascription. Mental states in turn are understood through folk psychology. This can invite an objection that such an approach to mental states presumes the description theory of reference for mental terms. But if one adopts another approach to mental terms, they can be taken to refer to natural kinds that are investigated by the neural sciences and having states belonging to such natural kinds cannot be a matter of ascription. Such an objection has been presented to eliminativism, but in some form it can be applied to the ascription theory as well. In the second part of this chapter, I argue that this objection is misguided and that the "flight to reference" fails to deliver its promise. In the final section of this chapter, I clarify the minimal requirements of the ascription theory for folk psychology, and rebut some objections to folk psychology thus understood.

4.1 Common-sense functionalism and the meaning of mental terms

The concept of mind and mental state terms are part of the common-sense framework that we use in our everyday life to make sense of each other. This is the basic use of mental terms; any other use of mental vocabulary (in cognitive psychology or in the neural sciences) derives from folk psychology. The idea that mental states like beliefs and desires are states featuring in folk psychology or that they are

central to our common-sense self-understanding is quite widespread. It is shared by the majority of those philosophers who take folk psychology to be a theory (so-called theory-theorists), common-sense functionalists and eliminativists like Churchland (1981) and Stich (1983), but also by philosophers of a non-eliminativist persuasion, who refuse to view folk psychology as a theory, but still attempt to understand beliefs and desires in terms of our common-sense conception (e.g., Baker 1995; Hornsby 1997).

The view that mental state terms are constituted by common-sense psychology can be supported with the following argument. One can partake in the activity of understanding oneself and interpreting others in mental terms without having any knowledge about brain states or other subpersonal processes. Understanding minds does not require understanding brains. Therefore, mental terms are to be applied on the basis of some other body of knowledge. The most obvious alternative is the common-sense conception of mind.

It may be objected to this that even if the understanding of mentality is relatively insulated, it does not follow that folk psychology captures all facts about mental states. I agree that one can come up with interesting stories about how various personal-level abilities are underpinned, but these stories are only initially provoked by our common-sense notions. Eventually, the theories of the cognitive and neural sciences use various successor concepts of folk notions. Indeed, one may say that through these studies we gain more knowledge about the mind, but this can be said only after one has picked up the concept of mind and other mental notions from its common-sense context.

However, in the argument sketched above, one can conclude something about the mind only when some premises are added. First, we should make the assumption, quite common in analytic philosophy, that minds and mental states are to be investigated through mental terms and concepts.[1] This leaves open the question of how such investigation would precisely proceed, but before one settles on the satisfactory form of analysis, one has to make the preliminary step and assume that such study would indeed inform us about the nature of mind and mental states. This assumption is in line with my treatment of mental entities as pleonastic entities whose nature is determined by the respective concepts. If this is granted, we may say that the study of mind is the study of the concept of mind. And if it is the case that the concept of mind is embedded in folk psychology, then one can conclude that the study of mind is a study of something folk psychological.

1. A recent example of a similar assumption is Millar (2004: 102): "There is no more to believing that summer has come to an end or to intending to learn more about European history than is specified by, respectively, the concept of believing that summer is coming to an end and the concept of intending to learn more about European history."

Second, one has to reject the idea that mind has a hidden nature that can be discovered by a posteriori investigations. It is a fact that we do not need to know neural science to understand everyday mental attributions. A natural conclusion from this is that the nature of the mental is determined by our common-sense psychology. But this is not concluded when one assumes that mind has an underlying nature. In this case, one cannot infer from a requirement on knowledge to the nature of a thing. Any such inference would seem to commit a kind of intentional fallacy, moving from some aspect of our knowledge of *x* to the nature of *x*. On the other hand, if the mental has no deep nature apart from what is involved in our practice with mental concepts, then the account of such practice would also provide an account of the mental. In other words, the question is whether mind is a natural kind. My position is that mind is not a natural kind (see Section 4.2): its nature lies on the surface, being determined by our concept of mind.

Someone may concede that folk psychology plays the central role in our grasp of mental concepts, but read this as a claim that folk psychology determines the meaning of mental terms, whereas their reference is left open by our common-sense understanding. It should be pointed out that the view that the reference of mental terms is independent from folk psychology contains potential seeds of the error-theory about mental discourse. It opens up the possibility that in using mental terms, we have been picking out entities and processes we did not take ourselves to be referring to. It may turn out that while we have seemingly been talking about beliefs and desires, we have actually been talking about certain neural states instead. I think that the origin of the strangeness of such a view can be traced to the peculiar understanding of the role of reference, both in our discourse and in philosophical argument. This issue will be discussed in the next section of this chapter.

Alternatively, it has been argued that folk-psychological specifications do not determine the meaning of mental terms, but rather play a reference-fixing role (e.g., Braddon-Mitchell and Jackson 1996). Note that this differs from the claim that folk-psychological specifications give us the references of mental terms, where the reference is any entity that satisfies a given cluster of folk-psychological descriptions. The idea behind such reference-fix talk is that the reference of 'belief' is fixed to the state that is picked out by the folk notion of belief, whereas the nature of that state can only be found out empirically. For example, according to a version of empirical functionalism, or 'psychofunctionalism' as it is sometimes called, the state that underlies the folk role has a functional essence which is to be discovered by psychology (Braddon-Mitchell and Jackson 1996: 80–82 describe the view in such terms). This is a kind of role functionalism where having a mental state is constituted by having a respective functional role. The folk-psychological specification (Braddon-Mitchell and Jackson call

this a 'role' too) merely plays a part in identifying a further role, the functional role. The folk role does not constitute the functional role as in common-sense functionalism. In the view described above, the folk specification does not exhaust the nature of the mental. There is no such route from the premise that we do not need to know or understand subpersonal facts in order to understand mental facts, for this account rejects the premise. The mental has a deep functional nature, and, therefore, even if we only need to know folk psychology in order to use mental terms, an account of the nature of mental facts must uncover their functional essence.

This view has its attractions. However, I do not think that it is viable.[2] In the second chapter of this book, I objected to role functionalism on metaphysical grounds. Here, my problems with such a view stem from the role of reference it assumes in the present issue. In the second part of this chapter, I argue that we should not rely on the notion of reference in resolving the question concerning the nature of the mental. Such an attempt presumes that there is a uniquely correct theory of reference. Without this unargued presumption, it fails to provide support to any view – be it 'mental terms do not refer', or 'mental terms refer to neural kinds', or 'mental terms fix reference to functional roles'.

In what follows, I am going to discuss the meaning of mental terms. If one shares the view that the nature of mental states is determined by the meaning of mental terms, then one can take this as a study of the nature of the mental. I highlight some problems for a certain version of role functionalism – common-sense functionalism. The central idea of this view is that the functional roles that give the meaning of mental terms are to be specified in terms of folk psychology. In this view, folk psychology is a collection of principles involving mental terms. It is presumed that every mental term is defined by the connections it has to other terms of these principles. Functional roles are taken to be causal roles that mediate inputs and outputs, and hence those folk principles should include input and output conditions. Take the folk principle that if someone desires *p* and believes that in order to get *p*, one has to do *q*, then one does or attempts to do *q*. For example, if one wants to drink milk and believes that there is a milk bottle in the fridge, one goes to open the fridge to fetch milk. If this were the only principle governing the desire to drink milk, then we could specify its role as being a state which is such that when combined with a proper sort of belief, it leads to milk-seeking behaviour. The obvious question is that of which folk principles would

2. One problem with such a view is that it seems to turn disagreements between empirical psychologists over subpersonal theories into disputes over the references of mental terms and hence as no genuine disagreements at all. (Cf. Kim 1996: 109 for such a line of argument.)

enter into the specification of the folk role. David Lewis, who has developed such a view on folk roles, has suggested that these should be principles that are our common knowledge.[3]

> Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. … Add also all the platitudes to the effect that one mental state falls under another – 'toothache is a kind of pain', and the like. Perhaps there are platitudes of other forms as well. Include only platitudes which are common knowledge among us – everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that names of mental states derive their meaning from these platitudes. (Lewis 1972: 256)

In order to prevent a situation in which a few such platitudes are false, Lewis proposes the formation of a cluster of these platitudes: "a disjunction of all the conjunctions of *most* of them". This cluster is the folk theory that gives the meaning of mental terms. The folk notions can be specified in terms of each other without circularity, by using a Ramsey sentence according to the technique proposed by Lewis (1970). We can write down the folk theory as $\mathrm{T}(m_1,\ldots, m_n, o_1,\ldots, o_m)$, where $m_1,\ldots, m_n$ are mental terms and $o_1,\ldots, o_m$ are other terms included in the platitudes. If we replace all mental terms with variables and place the quantifiers, we obtain the Ramsey sentence for the theory, which says that the theory is realised: $(\exists x_1)\ldots(\exists x_n)[\mathrm{T}(x_1,\ldots, x_n, o_1,\ldots, o_m)]$. It says that there are certain states that are related as specified by the theory. Now we can state the meaning of any *m* term. For example, if x_i is the variable that replaces m_i, then m_i can be given the following specification

$$m_i =_{df} \text{any } x_i \text{ such that } [(\exists x_1)\ldots(\exists x_{i-1})(\exists x_{i+1})\ldots(\exists x_n)\mathrm{T}(x_1,\ldots, x_i,\ldots, x_n, o_1,\ldots, o_m)]$$

The term is defined by a description with which we can pick out whatever entity is the *i*-th member of the realisation of theory T. Note that Lewis (1970, 1972) required that the theory has a unique realisation; in that case, the definition would be formulated as a definite description. Later, Lewis (1994: 417) allowed that the theory could have multiple realisations; in that case, its terms would be ambiguous. He also allowed that the roles could be imperfectly realised.

3. Lewis himself is a proponent of realiser functionalism: he identifies pain with the realiser that fulfils the pain role and not with the role itself. But since realiser functionalism also needs to start from role specification, in the present context I follow the common trend of proceeding from Lewis's account of role specification, which is the most developed account on offer.

To get role functionalism, we have to say that the functional property is the property of having a realiser property and not that it is the realiser that occupies the folk role. Such functional definition of a role property can be given as follows (cf. also Block 1980: 174):

> $M_i =_{df}$ the property of being an y such that $[(\exists x_1)\ldots(\exists x_n)T(x_1,\ldots, x_n, o_1,\ldots, o_m)$ & y has x_j (for some j, $1 \leq j \geq n$)]

Note that the definition quantifies over realisations of mental states in both cases.

The idea of common-sense functionalism is thus that mental terms are definable through the network of connections between inputs, outputs and other mental terms and that this network is provided by folk psychology. Lewis regarded such connections as platitudes, and this may invite the objection that, in fact, we do not seem to have a explicit grasp of folk-psychological principles. It is a notorious fact that no-one has been able to come up with a satisfactory and generally accepted list of such principles.[4] However, this worry can be answered by pointing out that our grasp of folk psychology may well be implicit (Lewis 1994: 416). Although we cannot normally formulate the principles and interconnections that specify mental terms, we are able to recognise them in their application. But what is implicit can be made explicit in principle (Braddon-Mitchell and Jackson 1996: 156). In other words, unless common-sense functionalists give special arguments against such explication, they must presume that there are no principled reasons why folk psychology could not be articulated one day.

Nevertheless, there is another objection forthcoming if common-sense functionalism is taken as an account of the meaning of mental terms. One may grant that we can still make use of folk psychology if we grasp it implicitly. But if the definitions of mental terms contain large gaps, then we have not defined mental terms. This problem appears because there is no complete list of all the principles, interconnections and terms that would fill the empty slots in the theory: $T(m_1,\ldots, m_n, o_1,\ldots, o_m)$. As long as the specification of a mental term includes such gaps, one cannot say that it means anything specific enough. For example, the belief that *p* cannot be defined as the state, which in combination with belief that *q* and desire to *r*, leads to the behaviour of type *x*, for it could well be that it would lead to some other kind of behaviour as well, or if it were combined with other mental

4. Attempts have been made, though. See Smedslund (1997), which is a rare and controversial attempt by a psychologist to work out stipulative definitions of psychological terms. He believes that his definitions and axioms capture important aspects of our common-sense terms. However, his system does not discredit my argument against common-sense functionalism. Setting certain doubts aside about this approach, I think that his specifications can be used to explicate some folk-psychological relations involved in ascription.

states, it would cause completely different output. The problem is that if the slots were filled with different terms, the definition would be different, and there do not seem to be any principled restrictions as to which slot-fillers are allowed and which are to be rejected. Common-sense functionalists are happy to say that, e.g., beliefs are whatever folk psychology says they are. The problem is now that folk psychology does not yield the unique specifications for mental terms. There seem to be endless ways one can come to believe that *p* and, similarly, it seems pretty much open which behavioural output would follow from the fact that one believes that *p*. This leads one to doubt whether it is even logically possible to generate definitions for mental terms that are both unique and complete. This problem is especially severe for an extremely holistic version of common-sense functionalism, which says that all connections between mental states enter into their specifications. This leads to familiar problems related to theory change: say, what to do when some theoretical connections change and how to even find out that they have changed. But even if we narrow down the list of defining principles to those that are not seriously doubted, the problem remains. Following Jackson and Pettit (1993), we may call the collection of such principles 'commonplace psychology'. But there is no reason to expect that the list of commonplace psychology will become explicit and that its definitions will not contain gaps. One may hold that beliefs are whatever commonplace psychology says they are, but the problem is that commonplace psychology does not say enough to fill the slot either.

A version of this worry is presented by Schiffer (1987: 31–32). He points out that functional definitions include the perceptual input condition of some such form (add any object in place of 'red block'): "If there is a red block directly in front of *x* and …, then *x* will believe that there is a red block in front of *x*." (Schiffer 1987: 31.) Schiffer argues that there is no list of conditions that should be added in order to fill the gap that would be common knowledge, either explicit or implicit. That there is no explicit completion of the gap is uncontroversial enough, given that no such list has been provided. As concerns implicit knowledge, Schiffer claims that if any such condition is part of the folk theory that defines the concept of belief, then the theory must be shared by anyone who has the concept of belief. But he finds it plausible that there could be people (probably visually impaired) who would possess the concept of belief, while failing to make an (implicit) use of the theory incorporating such perceptual input conditions. Schiffer (1987: 32) concludes that if folk theory contains such input conditions, it cannot give the meaning of belief since there might be people who possess the concept of belief without possessing the folk theory, and if such perceptual input conditions are omitted from the folk theory, the resulting network of interconnections would be too vague to yield a definition for the notion of belief. A defender of common-sense functionalism may dispute Schiffer's assumption that one can truly possess

a concept without possessing the theory that defines the concept, but I shall not pursue this line here. I agree that Schiffer has brought out an important issue, which relates to the more general objection that I presented above. The fact that functionalists are unable to fill the gaps in their definitions leaves their specifications of the meaning of mental terms seriously incomplete.[5]

I conclude that there is a major problem facing the account that attempts to specify the meaning of mental vocabulary in terms of folk psychology. Although the ascription theory acknowledges the close contact between folk psychology and mental terms, it does not incorporate the functional account of folk roles and it is not susceptible to the outlined argument. However, occasionally, when trying to be concise, I say that mental states are constituted by folk psychology. This should not be taken to mean that folk psychology implicitly defines mental terms. The idea is rather the following: folk psychology provides principles and interconnections that play a role in the application of mental terms. Such terms have more or less fixed non-holistic common meanings, which do not depend on their location in the folk theory. I present this view more fully in the third section of this chapter, where I argue that although the ascription theory presumes some folk principles, it can escape from the objections I raised against common-sense functionalism.[6]

4.2 Natural kinds and the reference of mental terms

In this section, I discuss the role of reference in the discussion of the nature of the mental. I begin with an objection that has been attributed to William Lycan (1988) that a certain view on folk psychology presumes a forlorn account of reference and consider Lycan's alternative proposal. The upshot is that "the flight to

---

5. One might argue that the defining specifications should be extracted from some scientific theory. Such theories are generally explicit and contain a tractable number of conditions. But there are two immediate objections to such a move. Which theory should one choose? There is no uncontroversial and single theory in psychology that we can rely on to provide the specifications of functional roles. Further, if there were such a theory, we could extract a definition of its terms, but what reason is there to expect that these, no doubt technical terms capture our mental concepts?

6. An additional problem with functionalism that I do not have space to discuss here, but which has been widely covered in the literature, is that the functional or input-internal role-output structure is not sufficient to specify mental states. This structure does not do justice to the so-called qualitative aspects of mental states as well as the normative import of mental state possession. I discussed the latter in Section 3.4.1 and I attempt to incorporate qualitative experience in the ascriptionist approach in Chapters 8 and 9.

reference", as Bishop and Stich (1998) call this move, is not successful and that the issue is not resolved by reference to the notion of reference.

Much in the way of setting up the problem concerning the relation between mental terms and the theory of reference and the later criticism of this set-up originates from Stephen Stich. He used to incline towards eliminative materialism (see Stich 1983), but later, after reading William Lycan (1988), he began to think that the case for eliminativism relies on a hidden assumption concerning reference. Afterwards, as recorded in Stich (1996), he changed his mind about the role of reference in the arguments for and against eliminativism. Although I do not think that eliminativism has to be formulated in reference-theoretic terms, I think it is still instructive to explore this way of seeing the issue in order to preclude the possibility that someone could level a similar reference-theoretic objection at interpretivism.

The ascription theory and interpretivism in general approach the mind through folk psychology and draw metaphysical conclusions about the nature of the mental from the premises about its folk-psychological conception. One may object to this approach that it makes use of the forlorn description theory of reference. If we held the causal theory of reference, we would be able to say that folk-psychological terms like 'belief' and 'desire' refer to certain empirically discoverable kinds and we need not presume that folk-psychological characterisation determines the character of the mental. A similar objection has been directed at eliminativism and, to my knowledge, it has not been applied to interpretivism. However, insofar as some versions of interpretivism are committed to the claim that having a mental state amounts to nothing more than being interpretable in folk terms, such an objection can be pressed against them as well. My aim is to argue that it is misplaced to rely on the notion of reference in an argument against interpretivism or eliminativism. Since the issue is more clear-cut and has been more widely discussed in the case of eliminativism, I concentrate on the latter, although the same moral applies to any attempt to argue against interpretivism proceeding from the notion of reference.

Eliminativism is the view that our common-sense understanding of the mental constitutes a thoroughly mistaken theory and is a candidate for replacement by an advanced neuroscience (e.g., Churchland 1981). The eliminativist argument can be represented as follows (Stich 1996: 3–4): mental states are theoretical postulates of folk-psychological theory; folk psychology is a false theory; therefore, mental states do not exist. Provoked by Lycan (1988), Stich came to accept that this argument relies on the hidden premise that the theoretical terms of a false theory lack reference. Presumably this premise holds only if the description theory of reference holds. The description theory of reference is basically the Lewisian account of theoretical terms construed as an account of reference: it constructs

a description for each theoretical term that refers to any such entity that satisfies this description. Now, the objection to eliminativism is that it cannot be maintained if one assumes the causal theory, which does not permit us to draw this pessimistic conclusion. The passage that struck Stich is the following:

> I am entirely willing to give up fairly large chunks of our commonsensical or platitudinous theory of belief or of desire (or of almost anything else) and decide that we were just wrong about a lot of things, without drawing the inference that we are no longer talking about belief or desire. To put the matter crudely, I incline away from Lewis's Carnapian and/or Rylean cluster theory of the reference of theoretical terms and toward Putnam's (1975a) causal-historical theory…I think the ordinary word 'belief' (qua theoretical term of folk psychology) points dimly toward a natural kind that we have not fully grasped and that only mature psychology will reveal. I expect that 'belief' will turn out to refer to some kind of information-bearing inner state of a sentient being…, but the kind of state it refers to may have only a few of the properties usually attributed to beliefs by common sense. (Lycan 1988: 31–32)[7]

The causal theory of reference (Kripke 1980; Putnam 1975a) pictures the reference-relation in terms of a causal-historical link between the object and its name. The description does not play a central role in determining the reference. If Lycan's suggestion were on the right track, then the dispute between eliminativists and their opponents would be transformed into a dispute over the correct theory of reference.

However, Stich later reconsidered the whole issue. Namely, the problem is that if eliminativists can be blamed for not having provided a reason for preferring the description theory to alternative accounts of reference, then the same charge applies to their opponents as well, for they should face a similar need to substantiate the causal theory of reference. This objection is developed in Bishop and Stich (1998). They point out that the argument moves from a claim about the lack of reference to a claim that some class of entities does not exist. For example, it is argued from the claim that 'belief' does not refer to the claim that there are no beliefs. In their view, such a move presumes the principle (R) that "something is an F if and only if 'F__' refers to it" (Bishop and Stich 1998: 36). This principle is also presumed by Lycan's reasoning, if he moves from the premise that '__ is a belief' refers to *i*, which is an instance of "the information-bearing inner state of a sentient being", to the conclusion that *i* is a belief and hence that there are beliefs.

7. This is more of a side note in Lycan (1988), but as already noted, the passage had a strong influence on Stich (1996), who discusses the role of reference in the debate over eliminativism extensively. Bishop and Stich (1998) also proceed from a reading of Lycan as making a case against eliminativism from the reference-theoretic point of view.

According to Bishop and Stich, neither eliminativists nor Lycan have shown that their theory of reference makes the instances of the principle true. (R) holds trivially under the deflationary account of reference, but then "the flight to reference" strategy is not available (see Bishop and Stich 1998: 39–40, n. 6). The description theory and causal theory are, however, both substantive or inflationary accounts. In the inflationary case, one has to show that (R) holds in one's theory of reference and, furthermore, that the theory of reference is the only such theory where the principle holds. The latter requirement stems from the fact that different theories of reference yield different results with (R) and if one wants to draw conclusions about what exists on the basis of (R), one has to assume that there is only *one* such reference-relation that grounds the instances of (R). Otherwise one would be committed to inconsistencies concerning what there is. The upshot of the discussion in Bishop and Stich (1998) is that those who want to rely on the notion of reference in arguing for or against eliminativism need to work out a theory of reference and show that the theory is unique in the sense that it makes the instances of (R) true. Meanwhile, there is no obligation to take "the flight to reference" seriously.

Bishop and Stich (1998) do not fault the specification of eliminativist argument in referential terms. They only point out that any such specification contains an unsupported premise. I agree with their diagnosis that eliminativism cannot be rejected by "the flight to reference" strategy. However, I want to substantiate a further claim that the debate over eliminativism does not have to be presented as being dependent on a particular theory of reference at all. This becomes visible when we see that both theories of reference can be combined with the allegedly opposing views on the nature of the mental (Crane 1998). For example, the description theory is consistent with the claim that the term 'belief' has an inner core, specified in terms of a causal role, that remains standing even when most of folk psychology has been eliminated. Similarly, the causal theory of reference is consistent with the claim that we have been referring all along to something completely different to what we expected, for a term refers even when our beliefs related to the concept are mistaken. For example, Lycan's alternative suggestion does not differ much from eliminativism, since it embraces the prospect that the inner information-bearing states may turn out to be something completely different from what folk have assumed about beliefs. When Lycan says that there are beliefs, he assumes that there are certain internal states whose nature has not yet been discovered. The only difference between this view and eliminativism is that eliminativism does not take such internal states as referents of mental terms, but this difference is only verbal. It is common to both views that they regard common sense as radically mistaken about the nature of belief (cf. Schulz 2002: 188). This shows that the question concerning the right theory of reference is orthogonal to

the issue of eliminativism. If the supporter of the causal theory can come to eliminativist conclusions, and if the description theory is compatible with the claim that the state playing the belief role can be realised even when much of the role description is mistaken, then one cannot say that eliminativism has to presume the description theory of reference.

Furthermore, I think that eliminativism can be stated without bringing the notion of reference to bear at all. In order to support the claim that there are no demons, one does not have to defend any particular theory of reference. The same applies to beliefs. The eliminativist claims that beliefs are posited by folk psychology as entities having such and such properties and that there really are no such entities. This characterisation does not rely on the notion of reference. Although, as we have seen, one can construe such claims as involving some theory of reference, this move is open to objections and need not advance the dispute.

In this section, three claims were argued for. First, drawing on Bishop and Stich (1998), an argument was made against "the flight to reference" strategy in general. Second, it was pointed out that eliminativism does not depend on any particular theory of reference. My final claim was that eliminativism need not be construed as being dependent on the theory of reference at all. Similar arguments can be advanced with respect to interpretivism and the ascription theory as a specific version of interpretivism. Neither interpretivism nor the ascription theory relies on any theory of reference and the notion of reference is not required to formulate these positions. I have refrained from investigating the notion of reference in this book. Taking mental entities as pleonastic entities does not entail immediate consequences for an account of the reference of mental terms.

What is the relation between eliminativism and the ascription theory? One may distinguish between conceptual and ontological elimination. Conceptual elimination aims at the wholesale displacement of mental terms with some scientific vocabulary. Ontological elimination rejects mental entities as a separate ontological category. In this respect, all those positions that reject inflationary mental entities, including those forms of reductionism that leave no place for separate inflationary mental entities, are instances of ontological elimination. The ascription theory eliminates mental entities in this sense as well. However, this differs from eliminativism as it is traditionally understood. Traditional eliminativism goes further than ontological elimination, for it recommends conceptual elimination. It does not expect to find any relations between mental and physical terms or, in other words, between the respective pleonastic entities. Whereas reductionist approaches relate pleonastic mental and pleonastic physical entities in various ways, eliminativism would purport to get rid of pleonastic mental entities altogether. For example, Churchland (1981: 86–87) envisages various prospects of the successor vocabulary of folk terminology. In addition, eliminativism is highly

critical of the theoretical virtues of folk psychology, whereas neither common versions of reductionism nor the ascription theory adopt this attitude. Eliminativists are bound to reject folk psychology, since they load it with several strong and unrealistic commitments – for example, it is assumed that folk psychology is committed to the idea that we have some kind of internal mental language that consists of separate semantically evaluable states that are causally efficacious. If the vindication of such claims were indeed required for the truth of folk psychology, then its prospects would seem shaky indeed. However, since weaker construals of folk psychology are feasible and more plausible, it can be argued that the eliminativist conclusion is an artefact of the strong construal of folk psychology.[8] I shall not pursue this line of inquiry further here; in the next section I give some additional reasons why eliminativism is not tenable.

I conclude this section with a note on the issue of whether mental states are natural kinds. There are two questions that require separation. One question is whether mental states *are* natural kinds and the other is whether mental terms *refer* to natural kinds. Someone who holds that mental states are natural kinds may understand this precisely as a claim about the reference of mental terms, but the first claim can be maintained without taking a stance in the issue of reference. Since I have already argued that the question of reference is orthogonal to present issues, I shall concentrate on the question of whether mental states are natural kinds. This issue is relevant here, for if mental states were indeed natural kinds, then the interpretivist account would remain superficial and would be unable to penetrate to the deep nature of the mental. To prevent this objection, I present some reasons for the opposing viewpoint. In a fairly recent account of natural kinds, Brian Ellis (2001: 19–21, 2002: 26–27) has brought out several requirements on natural kinds. For the present issue, three requirements are especially important. Ellis points out that the differences between natural kinds are

a. *objective*: they are not observer-relative;
b. *categorical*: one kind does not transform gradually into some other kind and for any element from the kind, there is always a determinate answer to the question of its kind-membership;
c. *internal*: the differences between kinds originate from the internal features of their elements.

Mental states do not satisfy these requirements. Given the folk-psychological understanding of mental states, it is highly implausible that the difference between e.g., beliefs and desires is a categorical and objective difference that comes from

8. For example, Horgan (1993) defends a conception of folk psychology that is "austere" and thus less susceptible to eliminativist strategies.

the internal nature of individual beliefs and desires. One can claim that mental states have functional or neural natures, but even then the requirements might not be satisfied. It is required that these functional or neural kinds are categorically distinct, based on their intrinsic differences, and it is doubtful whether there are such kinds in the brain. Those kinds that correspond to mental states are not to be individuated anatomically at the layer of neurones and cells. The proper layer of organisation where one would look for natural kinds corresponding to beliefs and desires is some higher, neurofunctional layer of organisation, and it is hard to see how the requirements for natural kinds – objective and categorical distinctness, intrinsic differences – can be satisfied at that layer.

Another case against the view that mental states are natural kinds is presented by Dennett (1994: 535–537). He takes the idea that mind is a natural kind as "a bulwark of hysterical realism" and develops a novel argument against it. Given that mental terms denote natural kinds, one should be able to give a non-arbitrary account of what these kinds are. Since natural kinds can be part of a larger natural kind, giving such an account is not straightforward at all. The membership of the natural kind in the case of mental states can be determined by some fact of the matter, hidden from folk and underdetermined by the ordinary usage of mental terms. However, this option is hard to justify, since one would then need a principled basis to claim that folk have denoted one kind rather than another kind containing the first, but no such basis is available. Dennett offers the familiar example of Putnam's essences of water – H_2O and XYZ – and conjectures that they could be part of some further natural kind.

> How could H_2O and XYZ not be instances of some single natural kind, given that they are as interchangeable in the physical world as Putnam requires us to imagine? Suppose, then, that H_2O and XYZ are both instances of some broader natural kind, K. Which natural kind did the folk mean by their pre-scientific word 'water'? Since they lacked any scientific purchase on the difference between H_2O and K, and could not, ex hypothesi, distinguish them, there cannot have been grounds in their usage or understanding for favoring one over the other. (Dennett 1994: 536)

The remaining option is that kind-membership is determined by our interests or judgements after all. And this is just another way of saying that mental kinds have no real, hidden essence.

4.3 The role of folk psychology in the ascription theory

In this section, I outline the view of folk-psychological terms that is required by the ascription theory and which is not vulnerable to the objections directed at the common-sense functionalism that defines mental concepts in terms of their folk roles. I also argue that several common objections to folk psychology, which may indeed be effective against some stronger construal of folk psychology, do not apply to the conception that is presumed by the ascription theory.

In principle, the ascription theory is compatible with the view that folk roles specify the meanings of mental terms. In that case, it would say that when we ascribe mental states, we ascribe certain folk roles to the person. This would wed the ascriptionist theory of possession to a functionalist account of the nature of mental states. However, I think that such a marriage would make the ascription theory vulnerable to the objections presented in Section 4.1. For if folk roles fail to specify the meaning of mental terms, there is nothing to be ascribed. Therefore, I am looking for an account that would still connect mental terms and folk psychology, but in which the point of connection would be subtly different.

I hinted at the conception already at the end of the last chapter. The idea is that mental (and other) terms have a meaning that is acquired when the language is acquired. The fact that our terms have the meaning they do and that it is quite stable is presumed by the ascription theory. To explain how our terms get their meaning is a separate task which is not undertaken here. Such facts about the meaning of mental terms are taken as the starting point and the present task is to explain what role is played therein by folk psychology. I assume that the meanings of mental terms are fairly minimal and uncontroversial, like basic "dictionary entries". They do not comprise anything as large and as holistic as the term's role in the whole theory. For example, we may say that 'S believes that *p*' means something like 'S is in a position to acknowledge the commitment to *p*'. (One may plug in some other meaning here if she finds this proposal disputable.) Knowing the meaning of mental terms is part of our everyday and common-sense knowledge, comparable to knowing what 'light bulb' means. This does not entail, above all, that such meaning-specifications are unrevisable in any strong sense. Given this, folk psychology can be regarded as a framework that brings together all the mental terms of our language and connects them with each other (hence the locution 'folk-psychological terms'). There is thus an assumption that these terms are related in various stable ways. Some of these relations can be called common-sense principles, but they are rough generalisations rather than anything approximating to strict laws. Such interconnections are involved in the identification of which mental terms are to be applied to the subject. In other words, the folk-psychological principles are part of the application conditions of mental terms. Of course,

besides folk-psychological principles and relations that guide the application of mental terms, information about the concrete circumstances under interpretation is also required.

We thus get the following picture: folk psychology does not determine the meaning of mental terms, but it is relevant for the mental in two ways. First, it provides the set of interconnections that makes the ascription of mental states possible. Second, we may call the collection of mental terms and their interrelations 'folk psychology' and this enables us to make a concise reference to a large bundle of terms. The popular slogan that the application of mental terms involves knowledge of folk theory is thus compatible with the ascription theory, if it is taken with certain qualifications. This claim is usually put forward as the position of common-sense functionalism, but given the required qualifications, it does not conflict with the account developed here. Namely, knowledge of the theory need not amount to more than the availability of some common-sense interconnections that allow one to apply mental terms as appropriate.

The picture presented is more resistant to previously raised objections than functionalism. For if certain common-sense principles turn out to be incorrect, then it follows from common-sense functionalism that those mental terms of which these folk principles were constitutive lack meaning. In some versions of functionalism, this would lead to the conclusion that the folk theory as a whole is false and lacks the realiser. Thus, the functional definition of mental terms can lead directly to eliminativism. Given the picture I am recommending, no such conclusion follows. After all, since no folk principles are strict and exceptionless, they are not apt for evaluation as correct or incorrect. It is better to say that some folk principles are more appropriate in a given context than others and some interrelations between mental terms are more common than others. Those more appropriate and common principles would play a more central role in the application of mental terms. But if in some cases certain mental predicates cannot be applied or certain mental states cannot be ascribed on the basis of such principles, alternative relations expressed in folk-psychological vocabulary can still be found that the ascription could rely on.

Given my approach to folk psychology and mental terms, we can also shed new light on the eliminativist proposal. If the proposed account of mental terms is on the right track, then the elimination of mental vocabulary would be effected only in conjunction with the elimination of a large part of our ordinary language. The elimination would thus be a change with much greater repercussions than the mere rejection of a false theory or some stagnant research program. Perhaps this prospect would not really disturb eliminativists, for some of their wilder fantasies envisage "a new system of verbal communication entirely distinct from natural language" (Churchland 1981: 87), but such a prospect is too far-fetched and not

a reasonable constraint that an account of the mind should follow. Although it is contingent which principles make up folk psychology, they are part of our self-conception that provides us with basic concepts by which we understand others and ourselves and can only be eliminated with the cost of abandoning this self-conception and a large part of ordinary language.

What about the objection that folk principles are always gappy, that there are no completed formulations of such principles? As we saw in case of functionalism, this led to the result that no mental terms have specified meaning. This conclusion does not ensue in the present case. Mental terms have meaning as long as they are part of the language. However, if one were to embark on making explicit the conditions of their application, then we would see that they contain gaps. This is one of the reasons why the ascriber is ineliminable from my account and why I agree with those philosophers who claim that interpretation is uncodifiable. If folk principles were complete, with no gaps, and if they held without exceptions, then we could give a complete formalisation of the ascription process. In that case, the ascription of mental states would be a simple case of deductive reasoning from premises to a conclusion that is determined by the principles of folk psychology. However, this is not the case. Quite a lot can be made explicit about what is involved in the ascription, but a complete codification of the interpretive process is not forthcoming. Hence, there is a degree of competence that remains implicit and this is subsumed under the notion of an ascriber. Such explanation may invite certain circularity worries and this issue is tackled in Chapter 6 alongside other objections. As we will see in the next chapter, the gappiness of folk principles is counterbalanced by the availability of various ascription sources. It is reasonable to expect that even when, on the basis of a single source, e.g., behaviour and a bunch of folk principles, a definite ascription cannot be made, the availability of a number of sources with their principles and interconnections will give enough ground for the application of an appropriate mental term.

Note that this picture of folk psychology requires only the existence of some rough links that relate different types of mental states to each other (as well as to action and the environment). There is no need for a stronger requirement that folk psychology have the structure of a theory in any robust sense of 'theory'. In recent years, the view that folk psychology is a theory has received a considerable amount of criticism. The issue is complex and multi-sided (for a collection of classic papers see Davies and Stone 1995). At one extreme, there are theory-theorists who take our everyday knowledge of mind to consist in the application of a folk theory. This view is opposed by simulation theory, which rejects the idea that our understanding of the mental is theory-driven. There are also mixed versions of these views and some prefer to analyse folk psychology in terms of social routines and heuristics rather than picturing it as a system comprising a

fixed set of theoretical principles and terms. At first glance, especially given its stress on folk principles, it may look as if the ascription theory is most compatible with the theory-theory. However, the ascription theory is compatible with both theory-theory and simulation theory, if we take it that they strive to model the phenomenology of mindreading or to explicate its subpersonal structure. Both theory-theory and simulation theory present concrete proposals on how the ascription is effected: whether it is a matter of employing the tacit theory or running a simulation of another's thought processes using one's own cognitive system. The ascription theory, on the other hand, presents a general philosophical framework for what it is to have a mind. As such, it does not set strong constraints on the viable options for how one comes to know about the mental states of oneself or others. The ascription theory is compatible with various subpersonal accounts of the processes underlying the ascription. Neither is the ascription theory committed to any specific views concerning the phenomenology of interpretation. In particular, it does not have to claim that one consciously entertains folk-psychological principles and makes conscious inferences about other minds. However, some authors have employed theory-theory and simulation theory to formulate competing accounts of the meanings of mental terms. In some such readings, theory-theory represents an approach similar to common-sense functionalism, while some extreme versions of simulation theory extract the meanings of mental terms from first-person experience (Goldman 1993, 2002). These are commitments that the ascription theory does not share. The theory-theory account is susceptible to objections presented against functionalism and the phenomenological simulation view is singularly implausible. In particular, attempts to base the meaning of mental terms on the phenomenology of experience are bound to be vulnerable to familiar Wittgensteinian objection's to private language. (For a thorough critique of Goldman's account of concepts, see Child 2002).

In characterising folk psychology, I have written often about mental *states*. This is in part a matter of convenience, for I intended that my claims applied also to mental events and processes. However, recently, some philosophers have criticised the idea that folk psychology postulates mental entities. In the remaining part of the chapter, I discuss this critique and give reasons why I do not think that the ascription theorist should be unduly worried about it. Christopher Gauker (2003a, 2003b) and Adam Morton (2003a, 2003b) have attempted to interpret folk psychology without a commitment to the view that beliefs and desires are mental entities. The limits of this book preclude me from discussing their accounts in full, but I shall make some comments on Gauker's approach. He subscribes to interpretivism and analyses beliefs and desires in terms of what is involved in the practice of the attribution of beliefs and desires. His basic idea is that when we ascribe beliefs, we do not ascribe internal states to a person. Rather, we make an

assertion on this person's behalf. In Gauker's (2003a: 242) view, "an attribution of belief is an assertion on someone else's behalf" and "an attribution of desire is a command on someone else's behalf".[9] The view he rejects is the common position that beliefs and desires are "theoretical entities, postulated for the sake of a certain sort of explanation and prediction of behaviour" (Gauker 2003b: 216). His criticism of such a view has two main strands: he opposes the idea that mental states are internal entities, or that they are anything like 'states' at all, and he rejects the common assumption that beliefs and desires play a role in the prediction and explanation. The proper role of beliefs and desires is communicative, while communication is not information-exchange as it is usually taken to be. For Gauker, communication is a matter of fulfilling practical goals. As for folk psychology, his view is that our predictions are inductive and do not require the ascription of mental states. Alternatively, they can be based on the competence or skill of the person. Given this, we may give shallow common-sense explanations in terms of such competencies and inductive regularities.

Gauker's approach is certainly refreshing. Once you are an interpretivist and do not take beliefs and desires at face value, why should you not take a further step and reject the commitment to mental states completely and downplay their role in explanation and prediction? However, I do not think we should follow the radical path of Gauker, who comes to reject the notion of content and attempts to present a new kind of semantics based on what is appropriate to assert in a context. A critical analysis of his proposals is beyond the scope of the present book. I only note here that his account of the attribution of belief and desire seems to me too restrictive.

The way I have proposed understanding mental states does not carry the commitments that have lead Gauker to throw the baby out with the bathwater. The ascription of beliefs, in my account, does not involve the commitment to internal entities. There is no deeper reality to having a belief than having it ascribed in accordance with certain conditions. In this opposition to internal entities, the ascription theory actually joins Gauker and Morton. To say, as I have said, that mental terms are applicable to one need not entail that they are applicable in virtue of the terms corresponding to internal mental entities. Rather they are applicable in virtue of us sharing a folk-psychological conception that determines which sort of mental states are required in order for us to make sense of given behaviour in a given environment. The 'state'-talk is a mere *façon de parler*; its being so is most

---

9. Although different, Morton's (2003a: 286) account is congenial to this: "To ascribe a belief in p to S is to advise or sanction acting as if p were true, if S is a reliable source of information about topics such as p."

evident in my deflationary treatment of mental entities as pleonastic entities.[10] As concerns Gauker's claim that beliefs and desires have no role in explaining and predicting behaviour, this runs contrary to common practice. His alternative suggestion that the prediction of behaviour is based on induction may work in some stable contexts, but our behaviour in general is not regular enough to justify the application of inductive reasoning. Even if we reject the nomological-deductive account of folk-psychological explanation, it does not follow that we should reject the very idea of explaining in terms of beliefs and desires. We may understand the latter not as deduction from laws, but rather in terms of citing certain mental states in order to make sense of some behaviour in accordance with our shared folk conception. I thus see no reason to reject this picture of folk psychology, given that we understand it in the modest way outlined in this section.

To sum up: this chapter dealt with the role of folk psychology in the ascription theory. The basic idea is that there are folk principles and interconnections that govern the application of mental terms, but they do not determine their meaning holistically in the whole theory as common-sense functionalism suggests. The fact that folk principles contain gaps is a severe problem for functionalism, but it is not detrimental to my approach. I also rebutted some potential objections that might emerge from reflection on the reference of mental terms and clarified that the rather modest requirements of the ascription theory carry no such commitments from folk psychology that are sometimes subjected to devastating criticism.

---

10. Baker (2003: 185) avoids the commitment to beliefs as entities thus: "a belief is not an entity. The root idea of belief is of believing. Believing that snow is white is a property; the term 'belief' is just a nominalization of 'believing.'" I am in complete agreement with this.

CHAPTER 5

The ascription theory

5.1 Preliminaries

In this chapter, I outline the ascription theory, the view that the possession of a mental state consists in its being ascribable in accordance with certain conditions. The chapter begins with some preliminary remarks on the nature of the project undertaken here, allowing us to distinguish my account from some other related approaches. In the next sections, I present the basic elements of the theory and introduce the notion of the canonical ascription. Finally, I discuss the position of the theory in respect to indeterminacy and uncodifiability.

As has already pointed out several times in this book, the ascription theory is an account of what it is to have mental states.[1] As mental states have contents, this is also an account of content possession. Setting up the issue in these terms presumes a distinction between explaining meaning and explaining content possession. In my view, the theory of content possession presumes a theory of meaning and not the other way round. I do not attempt to reduce linguistic meaning to mental content as several authors do, assuming the following sequence of reduction: meaning is to be reduced to mental content, which is to be reduced to natural properties. If, on the contrary, meaning is not to be derived from and reduced to mental contents, then there is no impetus for seeking a separate account of mental contents. In my view, when we have already explained what is it to have states with mental contents, we do not need to give a separate account of how mental content comes to mean what it does or how contents hook up with the world. Such an account is already provided by a general theory of meaning. The idea is that the content of mental states is specified by using the terms of ordinary language. In general, then, setting certain idiosyncrasies aside, in an ascription such as 'John believes that there is a dog', the content of John's belief is specified by terms that already have a common meaning.

---

1. As a theory of what it is to have a mental state, the ascription theory should be distinguished from studies of causal attribution of personality traits in social psychology. In addition, it should not be confused with the approach of Olson and Kamawar (1999). By applying Davidson's interpretivism in the field of developmental psychology, they outline "the theory of ascriptions", which claims that the condition of having beliefs is the ability to ascribe beliefs to oneself and others. My ascription theory is neutral on this issue.

This creates the task of providing a substantial theory of meaning. My stance is that such an account is in principle within our reach. One plausible way how this account can proceed is to explain meaning by the norm-constituting common use of our terms. This would be a version of the use theory of meaning that fixes the meaning of our words through certain constitutive social roles. If we want a theory of meaning that is consistent with the ascription theory and can provide the required support for the latter, then there is an important constraint on any such theory. In this case, meaning should not be explicated in terms of mental phenomena. For if the possession of such mental phenomena were in turn explicated along interpretivist lines, the account would be circular. If these mental phenomena were to be accounted for in other, e.g., in functional terms, we would have a class of mental phenomena whose possession is not covered by interpretivism, and this would shake the grounds of the ascription theory. This constraint disqualifies any functional role theory of meaning that relies on a mental token's internal functional role. Similarly, any such account that reduces the term's use-property to some mental mechanisms or mental dispositions will not do for the purposes of the ascription theory.[2] The required account, on the other hand, would take those social roles that govern the use of our terms as external to our minds: they are thus abstract and objective patterns of usage.

As the meaning of those terms that specify the content of mental states is explained by a theory of meaning, which is separate from the present account, the same applies also to the meaning of mental state terms. As noted in the previous chapter, I assume that mental state terms like 'belief' and 'desire' have a common linguistic meaning that is not relative to a person to whom beliefs are ascribed. Here, I want to present some reasons for the claim that ascribability is not part of the meaning of mental terms. Assume that the opposite is the case and 'believing that *p*' means something like the following: 'Being in the position to acknowledge the commitment to *p* is ascribable to S'. This would be a case of interpretivism about the *nature* of belief – interpretability is implicated in the very concept of belief itself. On this approach, no separate account of possession is required. In view of this, the ascription theory is incompatible with the theory that includes ascribability to the meaning of mental terms. This point has little effect on those who do not subscribe to the ascription theory, but it is an important clarification in respect of my present project.

---

2. For example, Horwich's recent account explains a word's meaning in terms of its use and explicates use as underived acceptance of sentences containing the word. He gives a nonsemantic account of acceptance by explaining it as having a mentalese sentence in one's internal "belief box" (Horwich 2005: 41). Since his account "goes internal", it is of no help to the ascription theory which does not rely on ascription-independent internal mental events.

Imagine thus that ‘S believes that *p*’ means ‘Being in the position to acknowledge the commitment to *p* is ascribable to S’. If we plug this account of the nature of belief into the ascription theory, we get

> The possession condition for S believes that *p*: Being in the position to acknowledge the commitment to *p* is ascribable to S is ascribable to S

This does not make sense, for the second ‘ascribability’ is clearly redundant. Thus an ascription theory about possession cannot incorporate such an approach to mental terms.

The view that the meaning of mental terms involves ascribability is open to critique on two counts. First, it is not correct as a matter of ordinary usage. It is hard to take seriously the proposal that ascribability is part of the meaning of mental terms, for the fact that such commitments are ascribable does not play a part in most cases where we apply the term ‘believes’. Ascribability remains transparent when we concentrate on the nature of belief. It comes into play only when we pose the question as to what it is to have a state with such and such a nature. This is to say that it plays a role in an account of the *possession* of beliefs. A second problem with this view is that the ascribability should presumably enter into the meaning of all mental terms, for only in that case would interpretivism about the nature of the mental be constitutive. However, it is singularly implausible that all mental terms involve ascribability as a part of their meaning. An interpretivist who takes this approach simply makes their theory implausible for no good reason.

Let us embark on another issue. In a nutshell, the ascription theory is a version of intermediate interpretivism at the level of possession. There is an ineliminable element of interpretation in the possession of mental states but the interpretation is not arbitrary: the correctness or appropriateness of an interpretation is objective. The theory can be viewed as consisting of two parts. First, there is the general interpretivist thesis that connects the possession of mental states with their ascribability. Second, there is a particular account about what is involved in the ascription and how it proceeds. Obviously, one may espouse the general interpretivist line without sharing the details of the ascription theory. But some more specific account is required in any case. In the next section, I shall make a start by presenting my own proposal as to how the general interpretivist approach can be filled out. In the remaining part of this section, I discuss the general approach.

One may challenge the general interpretivist claim that the possession of mental states involves ascribability. Such a challenge can proceed by presenting an alternative account of mental states and by pointing out that in this account, ascription is not required. Let us investigate this more closely. I am going to argue

for the general interpretivist claim. First, I argue that some alternative accounts that do not connect the mind with ascribability are actually compatible with interpretivism or involve implicit commitment to it. Second, I give some reasons for the need for ascription.

Three such alternative accounts come to mind. For simplicity, I concentrate on theories of what it is to have a belief. First, there is functionalism, which says that having a belief is a matter of instantiating a functional role that plays the belief part in the theory that gives a constitutive account of mental states. The second alternative approach is the dispositional account of belief, which claims that having a belief that *p* is a matter of having certain (behavioural) dispositions related to *p* like assenting to *p*, when prompted, etc. The third account is Lynn Rudder Baker's practical realism that explains belief in counterfactual terms. In all these accounts, there seems no need for ascription. In the functionalist picture, one believes that *p* if one has an appropriate functional role; in the dispositional account, one requires an appropriate disposition; and in Baker's view, certain sets of counterfactuals have to be true of the person.

However, in my view, these accounts are compatible with the interpretivist approach in the sense that it is possible to plug an interpretivist component into these views or that it can be shown that they covertly involve interpretivist assumptions. The reason for this, one can argue, is that instantiating a functional role, having an appropriate disposition and the truth of relevant counterfactuals are a matter of interpretation. For example, whether someone has a state playing the functional role of belief is a matter of whether one's perceptual inputs and behavioural outputs are of the relevant kind that warrant the ascription of a state intervening between these inputs and outputs, given one's other mental states. And this in turn can be a matter of interpretation. It is hard to see how inputs and outputs can themselves make it the case that one has a respective belief that *p*. Of course, the widespread assumption is that the causal relations between the realisers of inputs, internal states and outputs are responsible for the functional role. Specifically, it is presumed that the instantiation of realisers makes it the case that the respective functional role is instantiated. However, this approach is vulnerable to various objections, already discussed in Chapter 2. Here, my main purpose is to point out that one could in principle countenance the functionalist account of the nature of belief in terms of functional role, but claim that the possession of this role is not determined independently of interpretation.

As already stated, the same point applies to dispositional accounts. I discuss the dispositional approach and Baker's practical realism together. Although Baker (2003: 195) objects to certain dispositional accounts that reduce dispositions to the categorical basis, she concedes that her account is also dispositional in the shallow sense that analyses dispositions in counterfactual terms. Baker's account

explicitly concerns the possession of belief and not the meaning of 'believes'. Since she does not think that one can provide an analysis of 'believes' in terms that do not presuppose the very notion, she aims to explicate "what is required for someone to have the property of believing that *p*" (Baker 2003: 185). In a nutshell, her claim is that one believes that *p* if and only if certain counterfactuals are true about the person.

> Whether a person *S* has a particular belief (individuated by a 'that' clause in its attribution) is determined by what *S* does, says, and thinks, and what *S would* do, say, and think in various circumstances, where 'what *S* would do' may itself be specified intentionally. So, whether '*S* believes that *p*' is true depends on there being relevant counterfactuals true of *S*. (Baker 1995: 154–155)

Even this account could contain an interpretivist component. An interpretivist may agree with Baker that having a belief depends on counterfactuals outlining one's possible behaviour and utterances, but he would claim that what S does, says and thinks depend on the interpretation. The relevant behaviour is not a bare physical movement: it comes under intentional description and would thus require interpretation as to what type of behaviour it is. The same applies to utterances that need to be interpreted. In sum, an interpretivist can hold that what counts as the relevant behaviour, what one means by an utterance, and what thoughts can be ascribed to one, are all fixed by interpretation. One can therefore hold that having a belief is determined by certain counterfactuals, but this hides the fact that these counterfactuals are themselves interpretive.

Baker (1995: 218) stresses that in her account, having a belief does not depend on observers. She spells this out by saying that the relevant counterfactuals are true about the person irrespective of whether anyone knows that they are true. It may seem that this conflicts with the interpretivist proposal, for in the latter case, having a belief seems to depend on the interpreter. This issue merits closer inspection. Baker distinguishes between mind-dependence and recognition-dependence. She argues that the former fails to draw an interesting metaphysical distinction vis-à-vis objectivity and reality. Baker claims that recognition-dependence provides the required distinguishing mark, at least in respect of objectivity. In her view, a property is recognition-independent if an object can have this property 'independently of anyone's awareness of it' (Baker 1995: 232). She takes objective matters to be independent from one's judgements, enabling a "distinction between being correct and seeming to be correct" (Baker 1995: 218, n. 26). In her account, recognition-independent facts (like the fact that *x* has *P*), are such that our opinions about them "stand subject to correction" (Baker 1995: 233). According to her practical realism, having a belief belongs to this class of recognition-independent matters.

As stated, this distinction does not seem to be easily reconcilable with the approach of intermediate interpretivism, which claims that having a belief is recognition-dependent but nevertheless objective. However, it is likely that the seeming disagreement ensues from different construals of the notion of recognition-dependence. As I used the notion in Chapter 3, an instantiation of a property is a recognition-dependent matter if it involves recognition and hence the recognisability (or, as in present case, interpretability) is ineliminable from an account of what it is to have the property. Baker seems to use two different notions of recognition-dependence. One notion looks similar to the one outlined in the previous passage: the dependence of a property's instantiation on attitudes or knowledge. The other notion involves the idea that one cannot correct judgements about recognition-dependent matters, since in the case of such judgements, there is no distinction between correctness and apparent correctness. It does not seem straightforward that what is recognition-dependent in the first sense is also *ipso facto* recognition-dependent in the second sense. However, I shall not press this point here. Instead, I shall argue that intermediate interpretivism (and the ascription theory especially) construes having a belief as recognition-independent according to both of Baker's specific notions of recognition-dependence and yet recognition-dependent in my sense of the term.

Let us take the first sense of 'recognition-dependence'. It is indeed true that there is dependence between possession and interpretation according to intermediate interpretivism. But this is dependence at a level distinct from that which Baker requires, for it is compatible with the fact that statements about one's mental states are true of one independently of our knowledge or awareness. If the belief that *p* is ascribable to S, then it need not be ascribable to S *in virtue of* some actual interpreter's knowledge or recognition that it is so ascribable. (By 'actual interpreter' I mean anyone who makes ascriptions in our world.) This can be further clarified if it is formulated in terms of the ascription-condition.[3] Let us say that for S to believe that *p* is for S to satisfy the ascription-condition for the belief that *p*. Now, *the fact that* S satisfies the ascription-condition for the belief that *p* is independent from the knowledge of any actual interpreter. In particular, it does not obtain in virtue of any actual interpreter knowing that S satisfies the ascription-condition for the belief that *p*. According to the approach developed in this chapter, what is required is that the belief that *p* would be ascribable to S, but there is no requirement that the ascription should actually be made. In this respect,

3. Expressing the claim in terms of ascription-conditions need not entail that there is only one ascription-condition for every substitution instance of 'belief that *p*'. It means only that certain conditions are satisfied that make the belief that *p* ascribable. In the case of different persons, these conditions would differ, but the belief with the same content would still be ascribable.

Baker's requirement that one has the property of believing that *p* independently of anyone's awareness is thus satisfied, for the interpreter need not be an actual person. However, when we explicate what is involved in the satisfaction of the ascription-condition, we must mention interpretability. It may be objected that the satisfaction of the ascription-condition is still dependent on its recognition by the possible interpreter. I countenance this sort of recognition-dependence, but this dependence does not conflict with Baker's analysis, for she did not make (at least explicitly) the stronger claim that the instantiation of a recognition-independent property *P* by some *x* entails that it is possible that *P* is instantiated by *x* and that in all possible worlds, there is nobody aware that *x* has the property. However, if she intended this stronger construal, then the ascription theory regards having a belief as recognition-dependent in Baker's first sense. Nevertheless, I do not think that this would ruin the main point that Baker's account can be made compatible with interpretivism, for it would still be recognition-independent in the Baker's second sense.

Setting the actuality issue aside, the sense in which interpretation is involved in possession may be clarified with an example. Baker's own example of a recognition-independent property is someone's having committed a crime in such a way that nobody has realised that such an act has been committed. I think that this kind of recognition-independence of crimes is compatible with an analysis of what it is to commit a crime in terms of breaking the law, whereas the explanation of why the law holds is given in recognition-dependent terms. Having a belief involves recognition in this comparatively modest sense.

The second mark of recognition-dependence was the applicability of the is/seems distinction. As we will see later in the chapter, there is a distinction between the correct and the incorrect ascription on my approach. Baker claims that

> Ordinary human attributors of belief may be straightforwardly mistaken – either in identifying relevant counterfactuals or in assessing their truth value. So, belief is observer-independent. (Baker 1995: 218)

Similarly, the ascription theory claims that an ordinary attributor may be mistaken when ascribing beliefs, and that there is a criterion of correct ascription (see Section 5.4). Since ordinary ascribers need not always be up to the level at which the correct ascription is located, the attempts of ordinary ascribers are subject to correction. It follows from this that Baker's contention that believing is recognition-independent does not present an obstacle to the claim that her account of believing in terms of counterfactuals is compatible with interpretivism.

That is but a first step in support of the general interpretivist claim. The fact that some competing accounts may involve interpretivist elements does not show that interpretivism should be preferred to these accounts. What is needed is a

consideration that supports interpretivism. One way to present such a consideration is to show that the alternative accounts are prone to objections on independent grounds. Some such arguments were presented previously, especially in Chapters 2 and 4. However, even if one could reject all alternative accounts, this would not show that interpretivism must be true. We need rather an argument for interpretivism. An attempt to provide such an argument and to explicate its background assumptions is made in Chapter 6. Here, I just present the preliminary idea. It is again connected with the point that the mental pattern exists only in connection with those who can appreciate it as a mental pattern. If the pattern is indeed recognition-dependent (in my sense and, as we saw, not necessarily in Baker's sense), then some kind of approach that respects this sort of recognition-dependence is required. Interpretivism would thus be an obvious candidate. Of course, we still need to show that mental patterns are indeed recognisabilia. There are no explicit well-known arguments for this in the literature. I view it as a starting point of the interpretivist enterprise, as an assumption for which it is difficult to give an absolutely compelling argument. It is easier to assume simply that this is the case and to see how one can proceed, given such an assumption. Nonetheless, some considerations in support of this claim can still be voiced. Since the ascription of mental states and contents plays an important part in our common-sense interaction, it is reasonable to suppose that mutual recognisability is central to the possession of mental states. At least, the idea of in principle undiscoverable mental states does not make much sense, if we view mental states as embedded in folk psychology. However, this does not yet show that mental states are recognisabilia, for we might simply be well-placed to become aware of the recognition-independent beliefs and desires of our conversants. The additional impetus must come from the idea that mental states do not wear their labels on their sleeves. Various factors play a role in determining which state one has and which content the state has. A persuasive conjecture is that such factors involve and require interpretation. The claim is not that these factors themselves put labels on mental states; it is rather that the possession of a belief (and other mental states) has to be recognisable as the possession of a state of the respective type with a given content. As already indicated, this kind of recognisability does not preclude the application of the notion of correct interpretation to the mental case. In the view developed in this book, it is not an arbitrary matter which mental states one has. And now it is time to move on to presenting the promised account.

5.2 The ascription sources

In this section, a preliminary formulation of the ascription theory is provided. On this basis, the account is further refined in the following sections. The basic idea of the theory is that the possession condition for X's having a mental state *s* with content *p* is that the state *s* with content *p* is ascribable to X in accordance with certain nontrivial conditions. The view is thus a version of interpretivism, since it explicates the conditions of possession in terms of interpretability.

The notion of ascription is a technical term in my account. To ascribe mental states is just to relate a subject to a certain kind of a mental state (with a particular content). In terms of the metaphysical scheme outlined in Chapter 2, it is to relate the subject of ascription to a pleonastic mental property. No constraints are thereby placed on the various ways the ascription can be carried through. For example, it is not required that the ascriber makes the ascription with a fully articulated assertion in the vicinity of the subject. Neither does my modest theoretical use purport to capture all the intricacies of the ascription as a speech act or as a normative move in everyday communication. Thus, I do not attempt to make explicit what it is that we covertly do when we ascribe mental states, in the style of Brandom (1994) or Gauker (2003b). As has already been pointed out in several places in this book, the ascription theory does not purport to be an empirical or a phenomenological account of the actual behaviour of users of folk psychology either. No empirical commitments to the actual mechanisms of mindreading are made.

As concerns the actual phenomenon or the practice of understanding other people, this is quite messy. Not every interaction with others involves the ascription of mental states and when we do actually make ascriptions, sometimes the aim is not to gain a definite fix on the subject's mental states or on the contents of one's thoughts. Instead, the aim might be to blame others for what they have done, without taking a final stand on the issue of whether the other people really have the attributed intentions. Sometimes the ascription of mental states to others serves as a conversation opener; sometimes it is a device for claiming something on behalf of the other, as Gauker (2003b) suggested. Also, when one explicitly ascribes thoughts to oneself, the purpose might be to begin reasoning about one's own motives and intentions in order to find out what one really is or should be thinking instead of simply taking for granted what one's thoughts are (see Moran 2001 for an extended discussion of such matters). All of this is orthogonal to the point of the ascription theory, which analyses the possession of mental states in terms of their ascribability, which is an abstract and theoretical notion that leaves out all covert and overt intentions that people who actually ascribe mental states might have.

At the same time, the ascription theory seeks to keep in contact with ordinary practice. There are three areas of such contact. First, although more comprehensive evidence about the subject is presumed than we typically have, no such evidence is required for the ascription that could be uncovered only in very rare or artificial circumstances (like brain scans, hypnosis or round-the-clock surveillance).[4] Second, the way of coming to an ascription from this evidence does not require superhuman abilities. Third, it is assumed that the ascriptions are done in terms of folk psychology which is the common medium by which we understand mental states. This kind of contact with ordinary practice is required, for, after all, the ascription theory attempts to explain the possession conditions of the very same kind of states that we ascribe to ourselves and other people. If the ascription conditions by which the theory explains the possession of mental states are made so stringent that ordinary people are even in principle unable to meet them, then it would be a theory about some states other than mental states.

The ascription of a mental state is not arbitrary. It proceeds on the basis of various sources of evidence and is guided by the latter. The intuitive idea of the ascription theory is that these different sources play a part in determining which mental states one has. Such mental states are regarded as ascribable that enable us to make the best sense of the subject in a given setting. This requires that the ascribable states must be in harmony or cohere with the evidence from the ascription sources. The reason behind the requirement of harmony is the commonsense idea that the subject's mental profile is related to its environmental inputs and behavioural outputs and other ascription sources. I think it is safe to assume (with e.g., Levin 1988: 206) that according to folk psychology, our beliefs depend in part on what we see, our desires depend on what we require and what we aim for, our propositional attitudes give rise to behaviour, etc. Every relevant change in the ascription sources would warrant respective modifications in what can be ascribed. I presume thus that such interconnections are implied in folk psychology and that the modifications follow folk principles and connections.

What are the ascription sources? There are four main kinds of evidence that can be relevant in guiding the ascription. Note that every type of evidence is naturalistic, that is, they are publicly accessible and objective evidence about the subject. These types of sources are presented here in no particular order. I have also

4. In this regard, I agree with Child (1994: 25) that the ways of getting evidence for interpretation should not exceed the ways people ordinarily find out about each other. Contrast this with Byrne's (1998: 207) assumption that the Dennettian interpreter is "aided in her task by hidden high resolution cameras, shooting you from all angles; the Interpreter then views the resulting films in the comfort of her own home, aided by her compendious memory of what she saw herself."

pointed out some candidates for ascription principles for each source. I do not presume that these are rules that every ascription has to follow or that they are the sole conceivable candidates. Rather, the reason why they are mentioned here is to demonstrate in the case of each source that there are relevant ascription principles that relate this type of evidence to the subject's state of mind. To the extent that these principles can be viewed as the generalisations of various folk-psychological principles, reformulated as rules for ascription, mentioning them here also helps to motivate the thought that ascription relies on the folk psychological network that covers the relations between mental states, action, environmental conditions and other factors.

The first source of data is *other mental states with contents that are ascribable*. This involves both those relevant mental states that were ascribable in previous times and those mental states that are ascribable jointly with the current ascription. It goes without saying that all such mental states come under consideration with respect to their contents, not just their state-types. The importance of this source comes from the assumption that one's mental states are interrelated. For this reason, the fact that such and such propositional attitudes (and other mental states) were ascribable to the subject in previous times, and the fact that such and such propositional attitudes are additionally ascribable to the subject, influence the outcome of the ascription. Considering this source in the ascription of mental states respects the principle of holism: ascribe a mental state with a content by taking into account the subject's other mental states and contents. Of course, for different types of mental states, some other kinds of mental states are more relevant than others. For example, let us presume with Smedslund (1993) that person-directed anger consists in the belief that the person in question has unforgivably treated one disrespectfully. On the basis of this, it can be said that in order to ascribe anger towards Q to someone S, there must be a reason to ascribe to S the belief that Q has acted disrespectfully towards S and a reason to ascribe to S the state of not having forgiven Q's behaviour.

The second source of data is a subject's *behaviour*. This source is relevant, since there is a major common-sense assumption that one's behaviour coheres with one's mental states. From the interpretivist viewpoint, this is the requirement that one's behaviour must make sense in light of one's total ascribable mental profile. This differs from David Lewis's (1974: 337) "rationalisation principle", which says that the propositional attitudes ascribable to a subject should supply good reasons for her (physically described) behaviour. Lewis's principle presumes that subjects of interpretation are rational and conform to the principles of decision theory. Heuristically, this can be seen as a reasonable supposition, but since we want to say that irrational subjects (like those described by Levin (1988) and discussed in Section 3.4.1) still have mental states, we cannot make the possession

of mental states dependent on interpretation that proceeds from strong rationality assumptions. Another question which is important in respect of the present source is whether behavioural evidence consists of movements in the physical sense or is already under intentional description. If the interpretation were radical, then this source would indeed be taken to consist in bare physical movements. Then the resulting model would involve two steps: first, the interpreter observes some physical movement and then she applies intentional descriptions to such physical patterns. No initial data under the intentional specification is allowed.[5] But as already argued in Chapter 3, the ascription theory does not proceed from the radical starting-point and other ascription sources also contain intentional elements. Hence, no radical restrictions should be placed in here as well.

The third source is a subject's *environment*, or more exactly, the environmental *stimuli* that have an effect on the subject. This source is motivated by the requirement to maintain harmony between one's environment and the contents that can be ascribed to one's mental states. This follows the folk-psychological platitude that our perceptions and certain beliefs are caused by objects in our surroundings. This platitude supports a version of the principle of charity: ascribe beliefs that are true with respect to the world. It is true that in a given environmental setting, there is a long list of potential objects that may have an effect on the subject. In general, it is thus difficult if not impossible to construe the ascribable state and content solely from the objects in one's environment. For this reason, this source is usually not the primary source of ascription, but in conjunction with other sources it is still important in yielding the right sort of mental profile. Note how this differs from Davidson's account that places the principle of charity in the centre. In the context of radical interpretation, one has no other option than to start from relations between the utterance and the environment. One is then faced with a problem of how to single out the right sort of environmental cause for the utterance in the absence of other information. The ascription theorist need not worry about this, since as a non-radical interpreter, he can proceed from a variety of sources.

The fourth and final source of ascription is a subject's *personal background*, which involves facts about the subject's personal history and dispositions, their language community and idiosyncratic language understanding. It also contains facts about the physical condition of the subject's body (like whether it has some

5. Burwood, Gilbert and Lennon (1998: 105–106) describe such a two-step interpretivist model. An account of this kind is especially prone to the problem that the interpreter may not know whether the source of the observed movements comes from the subject or is a kind of radio-controlled marionette. I discuss this worry in Chapter 6. If we proceed from genuine behaviour, this worry can be overcome.

defects that may hinder the display of certain behaviour or taking in certain environmental stimuli). This source may look like a ragbag, but it cannot be denied that all such facts can be relevant in determining one's mental state type and fixing content to one's thoughts. To offer an example, let us see how a subject's language can play a role in belief ascription. Of course, beliefs can be ascribed for a variety of purposes, but the ascription which is definitive in respect of the belief's possession should specify the belief content only in terms whose possession is also ascribable to the subject, unless non-conceptual content is ascribed (see Chapter 8). In accord with Brandom's (1994) analysis, we may say that such ascription is *de dicto* ascription. According to Brandom, by *de dicto* ascription, one attributes to the subject the commitment to *p*, where *p* is specified in terms that the subject could acknowledge, without the ascriber himself taking responsibility for the expressions used in the specification *p*. An additional plausible requirement for the ascription that defines possession is that the content of the subject's mental states should be specified in the subject's own language. This, I presume, follows already from the stress on *de dicto* ascriptions. The well-known Kripkean puzzles of belief ascription strongly suggest that we cannot uncritically rely on the principles of disquotation and translation (see Kripke 1979).[6] I shall not discuss these cases and the more general problem of belief reports in this book, but my general contention is that several such puzzles would be resolved once we took into account the information from other sources and expand the range of principles that are relevant in the ascription.

These are the sources of information about the subject from which ascription proceeds. This information provides concrete content for the abstract interconnections and principles of folk psychology. In addition, successful ascription presumes that the ascriber is in possession of general knowledge of the world and logic, but these latter kinds of knowledge are not specific to the subject of ascription and may be kept fixed irrespective of the person under interpretation. However, such knowledge may be helpful in guiding ascription. For example, it makes sense to assume that a subject's thoughts follow the relations between objects or states of affairs (cf. Heal 1998: 486). When the state of affairs that *p* brings about the state of affairs that *q*, then given that the belief that *p* is ascribable to one, the belief that *q* is also normally ascribable. In this case, such a mirroring assumption would influence the outcome of the ascription through the first source

6. The principle of disquotation says that "If a normal English speaker, on reflection, sincerely assents to '*p*', then he believes that *p*" and the principle of translation says that "If a sentence of one language expresses a truth in that language, then any translation of it into any other language also expresses a truth (in that other language)". (Kripke 1979: 248–250, original italics omitted.)

above. Sometimes, the reasoning of some subjects may best be described by some highly idiosyncratic rules. In this case, these rules would influence the ascription through the fourth source above.

The ascription theory can be viewed as an elaboration of the suggestion, which I think is fairly common-sensical, that adequate ascription must take into account all available relevant information about the subject and make it fit together. More specifically, the claim is that one has a mental profile that would make the most sense to ascribe relative to the aforementioned ascription sources.[7] In order to sharpen this claim, let us assume that each source presents a declaration c_n for the ascription. By 'declaration' is meant a statement of the evidence presented by each source. For example, the second source would declare that the subject behaves in such and such a manner. Of course, one source may present more than one declaration, but for simplicity of exposition we may assume that we are dealing with a single declaration for each source. These declarations constrain the eventual ascription statements by folk-psychological interconnections and principles that are central to our self-understanding. Some of these principles were mentioned above in relation to each source. Let us take the specification of a mental state with a certain content as a statement *As*, which is chosen on the condition that it must be coherent with the declarations of the sources. In order to explicate how sources lead to the ascription, we need to differentiate between the sources, relative to a given ascription. Not all sources are equal in their impact: we need to add a measure of relevance *r* to each source. For example, in some situations, facts about one's background outweigh the behavioural source. That is to say, the fourth source that yields c_4 has higher relevance for the particular ascription than the second source. The declaration c_2 yielded by that source would not then constrain *As* in the way it would have done in cases where the source producing c_4 has low relevance. But in the present case, in light of the subject's background, his acts would receive a completely different explanation in mental terms.[8] Similar stories can be told about other sources as well.

We can see how irrational and false beliefs would be ascribed in such an account. I have discussed the ascription of irrational beliefs already in Section 3.4.1, so here I can keep my comments brief. If there is a sufficient amount of commonsense relations (either explicit or implicit) and relevant data so that declarations can lead to the ascription, then mental states are ascribable. It does not matter if

7. I have outlined the basics of this approach previously in Mölder (1998). Mölder (1997) stresses the requirement of coherence between one's mental states, behaviour and background information.

8. Information about one's background can trump even first-person reports. For some examples, see e.g., Doppelt (1978: 8) or Mölder (1997: 58).

the content of beliefs is mistaken about the worldly facts or if one's mental states are paradigms of irrationality. Here is an example of the ascription of a false belief about the world. Imagine that someone, call him George, prepares to visit the capital of Estonia. He announces this intention to his friends, makes plans for travel and orders his tickets. When his plane arrives at Tallinn Airport, George makes immediate arrangements to continue his journey to Tartu. When he finally arrives there, he emails his friends that he has reached his destination. Since Tallinn is the capital of Estonia, George was mistaken: in view of his actions, he can be ascribed the mistaken belief that Tartu is the capital of Estonia. Of course, this false belief may be manifested in other ways as well: imagine George taking a geography test where he has to match countries to their capitals and he matches Estonia with Tartu. Similar examples can be given about the misrepresentation of one's immediate environment. Such an ascription faults the principle that recommends the attribution of true beliefs, but as noted, if other sources need to be considered as well and if such an ascription coheres better with the declarations of the sources, the false belief becomes ascribable.

This presents an intuitive picture of how ascription would proceed. Another and much more complicated task is to explicate the central concepts. In the next section, I attempt to cast more light on the point that ascription must be harmonious with or cohere with the evidence in the form of the declarations of the sources. Evidential harmony or coherence is a notion that has no uncontroversial formal specification, although at the intuitive level it may be quite easy to grasp. After that, I develop the account further and explain how I understand the relation between ascription and mental events and what is meant by 'ascribable'.

5.3 Modelling the coherence

In what follows, I will explore how one can model the notion of coherence used by the ascription theory. Paul Thagard (2000), who works at the interface between philosophy and cognitive science, has stressed the importance of coherence in many areas of human cognition, including social cognition, and has crafted a prominent way of understanding coherence. Of course, there are different models on offer that can provide the required elaboration of this notion. For instance, the notion of coherence can be specified for the purposes of the ascription theory also in terms of one proposition being a reason for another (relative to the subject's doxastic state), as in Spohn's (1999) special coherence principle. In that case, the subject would be the ascriber and the propositions considered would be declarations of the sources and ascription candidates. Therefore, it is not claimed that the ascription theory must be committed to Thagard's notion of coherence. However,

Thagard's theory is a useful ally, as it has many applications and there are several algorithms for computing coherence that accord with his theory. It should be noted in addition that the ascription theory can be fitted into Thagard's approach in various ways. I have considered different options and I present here only the one that seems to me the simplest model. This section is quite technical, so those readers lacking patience with technical details and looking for the eventual formulation of the ascription theory can go straight to the next section.

Thagard's (2000: 16–20) basic idea is that coherence is constraint satisfaction. We have a set of elements which can stand in relations of coherence and incoherence to each other. These relations are defined for pairs of elements. If two elements are coherent with each other, then the relation between them is a positive constraint and if they are mutually incoherent, that is, if they cannot stand together, the constraint is negative. The elements will be divided into the accepted and the rejected ones. If both elements are accepted or both are rejected, then the positive constraint between the elements of the pair is satisfied. The negative constraint is satisfied if one element is rejected and the other is accepted. Thagard's claim is that assessing coherence can be viewed as solving the *coherence problem* – this is to satisfy the maximum number of constraints with preference for more important constraints. The importance of a constraint is measured by its weight. The coherence problem is thus to find the partition that maximises the total weight of the satisfied constraints. It may be the case that all constraints cannot be satisfied together. However, the aim is to find the partition in which the set of accepted elements, *A*, includes elements that have positive constraint between them and leaves out those that are negatively constrained. The measure of the sum of the weights of the satisfied constraints allows us to compare the coherence of alternative partitions. In this regard, Thagard (2000: 19) says that "coherence is maximized if there is no other partition that has greater total weight." For a more formal description of the model, it is best to quote Thagard himself:

> Let *E* be a finite set of elements $\{e_i\}$ and C be a set of constraints on E understood as a set $\{(e_i, e_j)\}$ of pairs of elements of *E*. *C* divides into *C*+, the positive constraints on *E*, and *C*-, the negative constraints on *E*. With each constraint is associated a number *w*, which is the weight (strength) of the constraint. The problem is to partition *E* into two sets, *A* and *R*, in a way that maximizes compliance with the following two *coherence conditions*:
>
> If (e_i, e_j) is in *C*+, then e_i is in *A* if and only if e_j is in *A*.
> If (e_i, e_j) is in *C*-, then e_i is in *A* if and only if e_j is in *R*.
>
> Let *W* be the weight of the partition, that is, the sum of the weights of the satisfied constraints. The coherence problem is then to partition *E* into *A* and *R* in a way that maximizes *W*. (Thagard 2000: 18)

Thagard (2000: 28) notes that it has been proved that the coherence maximisation problem is computationally intractable, for the time required for computation increases exponentially with the growth of the number of elements. However, this is not a problem for his account, for there are various algorithms, connectionist and otherwise, that can give approximate solutions to the coherence problem.

The general characterisation of coherence can be applied in various fields through particular coherence models. For each such application, one has to specify the nature of the elements and constraints as well as the algorithms for computing coherence. Thagard's account of coherence as constraint satisfaction can be also applied to the problem of explaining other people's behaviour. This application is discussed in Thagard (2000: 101–109) and in Thagard and Kunda (1998). In the explanation of others' behaviour, the elements are propositions depicting the behavioural evidence and competing hypotheses that explain this evidence. The aim is to find the proposition that explains the behaviour and to reject the alternative explanations. For this task, the constraints are taken from Thagard's more specific theory of explanatory coherence (see e.g., Thagard 1992). In this account, there is a positive constraint between propositions if one explains the other. In addition, simplicity is seen as a virtue: the more hypotheses we need to explain the evidence, the weaker are the positive constraints between the evidence and hypotheses. On the other hand, if two propositions provide independent and competing explanations, there is a negative constraint between them. Another source of negative constraints is logical contradiction. Let us take the example used by Thagard and Kunda (1998) and Thagard (2000: 104–105): the behavioural evidence to be explained is the fact that Mary, who is usually a good-natured friend, screams at you. There are several alternative hypotheses that can explain the behaviour. For instance, she might suffer from fits of anger, she might be stressed by her work, she could have been replaced with an identical-looking irate twin, or you have done something that upset her. Which one to choose? To solve the coherence problem, this situation can be represented as a network of propositions. On the assumption that the hypotheses are independent, they are negatively constrained. Since each of them explains the evidence, the constraint between the evidence and each hypothesis is positive. Which hypothesis will eventually be accepted depends on the strengths of the constraints in the network.

The problem of finding the ascription that makes most sense in light of the sources can be modelled as a coherence problem in Thagard's sense. For this purpose, let us take the set of elements as consisting of source declarations $c_1, \ldots, c_4$ and a finite number of alternative ascription statements $As_1, \ldots, As_n$ (or collections of statements, if we want to ascribe several mental states simultaneously). Since I prefer a philosophically less loaded term than 'proposition', let us call these

elements, declarations and ascriptions, 'statements'. To formulate the coherence problem, all we need is to set up constraints between these statements.

Thagard's general characterisation of coherence does not discriminate between the elements. When computing coherence, all elements are treated as equal in the model. In order to accommodate the assumption that sources have different degrees of relevance, we need to set up the discriminating coherence problem. To give priority to certain elements, Thagard (2000: 71) provides a simple amendment to his general model: "For each element *d* in the discriminated set *D*, construct a positive constraint between *d* and a special element e_s, that is assigned to the set *A* of accepted elements." The discriminated set is a set of favoured elements, that is, elements which tend to be accepted. If we create the strong positive constraint between the special element and our favoured elements, the result is that due to this positive constraint, the favoured elements are more likely to be assigned to the set of accepted elements as well.

Equipped with these tools, we can translate the relevance measure of each source as the weight of the constraint between the declaration of the source c_n and the special element e_s. By definition, e_s is always accepted, so it should be a statement that in all partitions belongs to the *A* set, and it should be such that there can be a positive constraint between that statement and the declarations of the source. It could be any logical truth, but in that case the question may arise as to how its content is related to a given problem set. Perhaps it is better, then, to take e_s as any indisputable statement about the given context of ascription, for example, 'X is the subject of the ascription', for any appropriate substitution-instance of X.

The higher the relevance of the source, the stronger the constraint is between c_n and e_s. The idea is that since the ascription has to consider all the relevant sources, the declarations of the relevant sources always tend to enter into the set of accepted elements. In some cases, the source can be irrelevant for a particular ascription. Then the source does not present a declaration and it will not be a part of the model. I cannot think of a need to assign a negative relevance to some source. However, when some source yields a declaration that contradicts or competes with the declarations of other sources, then there is a negative constraint between these declarations. In cases where it is important that the declarations of some sources either rise or fall together, a positive constraint can be added between them. In that case, these declarations are taken as mutually supporting evidence.

In addition to constraints between c_n and e_s, there are constraints between source declarations and alternative ascription candidate statements. The weight of these constraints is determined by the explanatory relations between the ascription statement and statements of the sources. Explanatory relations are taken here fairly broadly, thus including all sorts of ways of making subjects intelligible in mental terms by the light of the ascription sources. These relations are

undergirded by the folk-psychological principles and interconnections. Negative constraints are assigned by the occurrences of contradiction and competition between the statements. It is assumed that alternative ascription statements compete with each other, and so there are negative constraints between them. As in Thagard's model, the coherence problem is solved when we find the partition of elements that maximises the total weight *W* of the satisfied constraints.

An additional condition on the partition, which is specific to the present account, is that the subset *A* of accepted elements has to include one ascription statement. Since there are negative constraints between alternative ascriptions, it cannot be the case that the partition in which more than one ascription statement, e.g., As_1 and As_2 are contained in *A*, has the higher total weight *W* than the partition in which *A* contains only one ascription candidate. However, it might be that we get a partition with a higher total weight *W* when there are no ascription statements in *A*, but this will not do for our purposes. After all, the aim is to find out which ascription statement coheres with the ascription sources and, for this, we need to measure the coherence of a set that contains both the declarations of ascription sources, $c_1, \ldots, c_4$ and the ascription statement.

To illustrate the overall approach, the ascriptionist network is depicted in Figure 1.

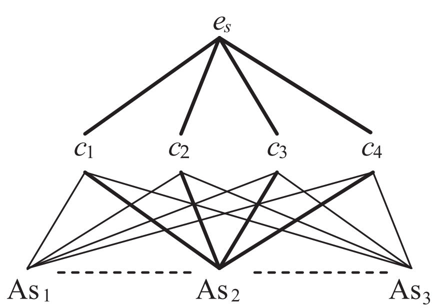


Figure 1. The ascriptionist coherence network

In the figure, solid lines represent positive constraints and negative constraints are shown with dashed lines. Stronger constraints are shown with thick lines. Potential constraints between source statements are omitted from the figure. We can see that there are strong positive constraints between the special element and source statements. There are negative associations between alternative ascription candidates. It is also indicated that strong positive constraints hold between the statements of sources and the ascription statement As_2. Hence it can be expected that the partition in which As_2 belongs to the subset *A* of accepted elements has higher *W* than alternative partitions in which either As_1 is in *A* or As_3 is in *A*.

In this model, the relevance of the sources is fixed in the sense that a given source has a certain measure of relevance which does not change during the ascription. The resulting ascription is thus relative to the assignment of relevance to the sources. If the relevance measure is fixed, we can compare different ascription statements with the aim of finding the one that maximises the coherence of the partition in contrast to other partitions in which the set *A* contains other ascription statements. If the relevance of (some of) the sources changes, the weight of the constraints between c_n and e_s changes, and this may influence the overall coherence of the partition. Ascriptions made on the background of different relevance assignments cannot be compared with each other, for they would be part of distinct instances of a coherence problem. This is one source of the variance in ascriptions in my account. I shall return to this issue in Section 5.5.

5.4 The canonical ascription

The explication of the ascription sources and the notion of coherence is but one aspect of the ascription theory. In addition, the relationship between ascription and mental events and the very notion of ascribability have to be clarified. If we explain the possession of mental states in terms of their ascribability, we need a criterion for the kind of ascription that the interpretivist account can utilise for its purposes. The reason for this is that not every particular ascription is up to the required level. Often we make ascriptions on the basis of insufficient data, or we do not pay attention to the different aspects of the situation. Thus, if the theory were to claim that X has all those mental states that one could ascribe to X, it would hardly be believable. What we need is to fix a standard for ascriptions that can play the constitutive role that is designated for them in the ascription theory. I shall call the ascription that conforms to this standard a *canonical* ascription. Using this term, we can say that one has a mental state provided that it is canonically ascribable to one. This notion can also be used to draw the distinction between the correct and the incorrect ascription. Hence the distinction is an interpretivist version of the is/seems difference. It need not coincide with the pragmatic notion of correctness: what is correct to say about someone's mental states in a given context is determined by conversational goals. The distinction between the correct and the incorrect ascription, in my sense, is the distinction between the canonical and the noncanonical ascription. This allows for the correction of ascriptions, for if a given ascription does not conform to the standard of canonicality, it can be corrected in order to make it canonical.

Even if we grant that canonicality provides the distinction between correct and incorrect ascriptions, someone may argue that one question remains

unanswered: why is the canonical ascription correct? Note that the ascription theorist cannot explain the correctness of the canonical ascription in terms of its capturing some antecedent mental facts. The point is rather that the canonical ascription supplies the condition that fixes the obtaining of such facts. Its correctness is to be explained not in terms of correspondence, but through the internal features of the ascription process, by reference to the manner in which the ascription is performed. The aim is to set such constraints on the ascription process that match up with folk-psychological intuitions about what is deemed important in mental state ascription. These intuitions are accessible to us when we reflect on those conditions that may lead us to revise the judgements concerning someone's mental states. I presume that the list of sources and the condition that the interpretation must cohere with the information from these sources, as outlined in the previous section, mirror these folk-psychological considerations. In addition, the canonicality condition is further elaborated in this section to make it compelling in light of certain philosophical concerns.

I claimed that the canonicality condition can be viewed as a standard of assessment for the acceptability of ascriptions. But most importantly, it provides us with the possession condition for mental phenomena. Note that it is not the case that there is a unique possession condition for each type of mental state and content. Mental states are not individuated by possession conditions. The possession condition is rather a general scheme that explicates the relation between mental phenomena and their possession and it applies to all mental phenomena. I am thus presenting an account of the following form ('X' stands for a subject of mentalistic ascription, '*f*-ing specifies a mental state and '*p*' gives its content; 'X *f*'s that *p*' thus describes a subject in mental terms[9]):

Possession condition for X *f*'s that *p*: *f*'s that *p* is canonically ascribable to X

This captures the sense in which it may be said that to have a belief that *p* is to have it canonically ascribable. It is presumed that all further conditions that an agent and its surroundings would have to satisfy in order to be suitable for a given interpretation are captured by the information contained in the ascription sources. Hence, a philosophical explanation of what it takes to believe that *p* may cite

9. I use 'that *p*' mainly for convenience. It should not be read as implying that contents *always* take some propositional form. For example, if there are mental states, which content is not propositional, there would be similar possession conditions for them. In this case, the ascription statement would take the form 'X *f*'s *p*', where '*p*' indicates non-propositional content ('Jane looks forward to her wedding') or 'X *f*'s', when there is no content to be specified ('Jane is depressed'). Similarly, the account is meant to comprise mental states like intending to ϕ, which specification does not use that-clause constructions.

various facts about the subject contained in the ascription sources, but according to the sort of interpretivism expounded here, such an explanation would not be complete without the claim that in view of such and such facts, such and such a mental state is canonically ascribable.

The ascription theory can be held at the level of predicates or at the level of properties or facts. We may take it as an account of the application conditions of mental predicates. In terms of the scheme outlined in Chapter 2, we may also say that the canonical ascription determines what it is for a subject to have a mental pleonastic property. Alternatively, it can be said that it determines the condition in which the mental pleonastic fact holds. Note that there are two facts. In addition to the pleonastic fact that X *f*'s that *p*, there is another pleonastic fact that *f*'s that *p* is canonically ascribable to X. It is not claimed that they are one and the same fact, since I individuate facts in a fine-grained manner. Those true sentences from which we can infer facts by something-from-nothing transformation have different meanings. These sentences cannot be replaced with each other in most contexts. There is thus no conceptual reduction of one type of facts to another type of facts. Therefore, the ascription theory is formulated as presenting a possession condition and not an account of the meaning of 'X *f*'s that *p*'.

Let us proceed to the clarification of canonicality. An arbitrary ascription will not do for our purposes. This leads to the thought that perhaps we can specify the canonicality condition through the notion of an ideal ascriber. We might say that one has such mental states that an ideal ascriber would ascribe to one. The ideal ascriber is omniscient with respect to the sources and can assess the coherence of the information better than anyone else. However, this account has a serious deficiency. Namely, it makes all of our ordinary ascriptions noncanonical, for ordinary folk cannot instantiate the ideal ascriber. It follows that we are always mistaken about others' mental states, and, what is worse, we are also mistaken when we engage in self-ascription. No one can be the ideal ascriber, even in respect of himself, for no one has privileged access to the ascription sources. Note that the possession condition is outlined in terms of what is ascribable rather than in terms of what is in fact ascribed. If we specify the canonicality condition through the ideal ascriber, the consequence from these two conditions is that in each case of actual ascription, there is some mental profile ascribable by the ideal ascriber which may differ from any profile that could actually have been ascribed.

Hence it is better not to fix the canonicality condition to the notion of an ideal ascriber. Since we routinely ascribe mental phenomena in everyday life, it would be astounding if the canonical ascription were in principle inaccessible to us. But if we presume that the canonical ascription is available to us, then this may seem to lead to another kind of worry. Presumably there is a method for the canonical ascription such that if we were to follow it, we would be canonical. This

leads to a *dilemma*: this method is either presently available or it is the method we would reach at the ideal end of the enquiry.[10] If the former is the case, then it is legitimate to ask why we should assume that this method cannot improve. But if the method can improve, so can the standard of canonicality and it follows that ascriptions done in accord with the present method fail to be canonical. If we reach the canonical method at the end of inquiry, then we are faced with the consequence that we presently have no idea what mental states we now "really" have. I have already argued that the "ideal" horn of the dilemma is not an acceptable option. What about the first horn? I shall argue that the ascription theory is not committed to this either. On the one hand, although one may say that there is a method of canonical ascription, the method is not of the kind that is detailed enough in order to be subject to subsequent revision and improvement. This has to do with the uncodifiability of the mental, which is further explained in Section 5.5. On the other hand, those constraints that the canonicality condition places on ascription relate to the fundamental aspects of our common-sense understanding of the mental. The idea of the sources and the notion of coherence are so central to our understanding of the possession of mental states that there is no room for improvement here. If we come to reject the idea that one's mental states help to make sense of one's behaviour and that what you believe depends on what else you believe, etc., then we have revised our very concept of mind. In addition, as we will see, the standard of canonicality as it will eventually be spelled out is quite relaxed. In view of that, even if one could make sense of the improvement of the canonical method, this would not have an effect on the results of the ascription.

Let us proceed with the idea that the canonical ascription is available to us. We need to locate the canonicality in the space between an arbitrary and an ideal ascription. It is not enough to say that the canonical ascription is the one that takes all relevant sources into account and finds the ascription statement that coheres best with the information provided by the sources. The reason for this is that such a standard is unable to supply the required distinction from the ideal ascription. If we presume that it is possible to approximate the maximum coherence, then for any ascription one can actually make, there will always be a more coherent ideal ascription. Thus we are back on the path that has already been rejected.

I suggest adding a relaxation clause to the canonicality condition. Let us start from the idea that any actual ascription, which is such that there is no need to revise it, would be suitable to determine a subject's mental states. Note that the possession condition is specified in terms of what is ascribable to one. It is not

10. This dilemma was presented to a previous version of the ascription theory by Wolfgang Spohn. An analogous dilemma that applies to physicalism is known as Hempel's dilemma, after Hempel (1980).

required that the ascription must be made in fact, but it is required that the ascription can actually be made. In saying that the canonical ascription must be such that it can be actual, I mean that it should rely on the sort of information and involve the sort of computational processes that are in principle accessible to ordinary folk. We *can* be canonical ascribers according to this standard of canonicality. This implies that canonical ascriptions are to be found in a world like ours, and the threat from omniscient ascribers is thereby blocked. The condition concerning revision is included to convey the idea that any ascription that coheres with the sources and that is acceptable when ascribed, is up to the standard that fixes the possession condition. An ascription is acceptable in a given situation if there are no warranted objections or actual defeaters to it. As any ascription is done in a particular setting, the features of that situation determine what sort of information is relevant and which revisions may be needed. Note that it is not required that the subject of ascription herself needs to agree with the ascription. There is no absolute first-person authority. However, such protests from the subject are usually a good sign that the ascription may need revision. Perhaps not all sources were considered or the relevance assignment to the sources needs to be modified.

It may seem strange that the canonicality condition involves unrevisability, for why should we assume after all that there are any unrevisable statements. However, the unrevisability in my condition is benign. It does not mean that we cannot revise our ascription. This is not excluded. The idea is rather that once we have reached an interpretation that does not invite any further revisions, the required interpretation of a subject's state of mind is settled. It does not make sense to ask whether, if some new evidence becomes available and we are compelled to revise our original interpretation, this means that the original ascription was not canonical, contrary to what we assumed, or that we have revised the canonical ascription after all. By definition, in the hypothetical situation in which the canonical ascription is made, all the relevant evidence is on the table, so there cannot be any new data that could overturn the original ascription, if it was canonical. Moreover, since the situation is hypothetical, it does not make sense to talk about the prospects of subsequent revisions. There is no "earlier" and "later" in relation to this abstract condition. Note that we do not need to give a separate account of the identity conditions of ascription situations. The theory proposes understanding the canonical ascription in such a way that in the situation in which the ascription is actually made, it does not require revision, and this claim already circumscribes the situation finely enough.

A consequence of this account is that it can happen that we do not recognise the canonical ascription when it is actually made. The reason for this is that although we can in principle instantiate canonical ascribers, we need not always be

sufficiently well-placed in respect of the sources and the coherence condition. We do not know whether we have taken all relevant information into account. But this characterises our epistemic condition in general and it is not a deficiency in the present account, which does not purport to compile a practical vade-mecum for everyday ascribers. Instead, it attempts to provide a metaphysical condition for the possession of mental states. However, this consideration leads to the following point: we need to distinguish between considering the ascription as canonical and the ascription being canonical. An ascription that we consider canonical need not be canonical. For this reason, the revisability constraint is not to be phrased as 'if the ascription were in fact made, it would not *be revised*'. We need, rather, a clause along the following lines: 'if the ascription were in fact made, it would not *require revision*'. To see the need for this, consider a situation in which we have two alternative ascriptions but we are unable to decide which one is better and therefore neither of them is subjected to revision. (Assume that if one of them were canonical, then the other would required revision.) In such a situation, we do not have two alternative canonical ascriptions. Rather, we should say that one of the ascriptions requires revision, but this is not undertaken. Similarly, if the ascription is left unrevised for some other reason, this does not entail that the ascription is canonical. Therefore, we need to put the clause in terms of not requiring the revision.

Observe that, given this relaxed formulation, the range of ascriptions that are coherent and do not require revision can be quite large. In my account, they are all canonical ascriptions. In a sense, this allows us to ease the situation of an interpreter who does not know whether he is canonical or not. Given that the ascription aims at coherence and that there do not seem to be any objections forthcoming, one can expect that one's ascription fits into the range of canonical ascriptions, although no absolute guarantee for this can be provided.

After all these remarks, we are finally in the position to fill some gaps in the formulation of the ascription theory. Let us take the following formulation:

> Possession condition for X *f*'s that *p*: *f*'s that *p* is canonically ascribable to X

Let us call 'X *f*'s that *p*' an ascription statement, *As*. Now we can state what it is for an ascription to be canonical:

> *As* is a *canonical ascription statement* $=_{\text{df}}$ *As* is a statement in approximation to the maximum coherence with the ascription sources $c_1,\ldots, c_4$ and if *As* were in fact ascribed, it would not require revision.

What we still need to explicate is the notion of a statement's being ascribable. Before we come to this issue, however, a few remarks are in order. Note that there cannot ensue any conflict between the two clauses of the canonicality condition.

On the one hand, if the ascription coheres with the available information, there cannot arise any need for revision. On the other hand, if revision is required, then there were presumably deficiencies in the appraisal of sources or coherence.

The second clause is not redundant, for, as we saw, it is needed in order to relax the high standards of the coherence condition, for the most coherent ascription statement, given the sources, may not be accessible. As I shall argue in Section 5.5, it also provides the requisite protection from full-blown indeterminacy and uncodifiability. The second clause can be compared to Dennett's predictivity condition. Recall that Dennett's (1987: 15) view is that any "system whose behavior is reliably and voluminously predictable" from the intentional stance is an intentional system. Note that Dennett does not really clarify which conditions a prediction has to satisfy in order to be "reliable" or "voluminous", but the idea is intuitive enough. A prediction is correct is the predicted event really occurs, and the method of interpretation is reliable if it keeps yielding correct predictions.[11] Predictivity as a standard for acceptable interpretation is quite coarse-grained. A given belief can be part of different interpretations which all yield correct predictions of the ensuing action. Similarly, my clause about the ascription not requiring revision allows for a broad range of ascriptions that may otherwise differ over minute details. If one assumes that the ascription is correct only if it matches up with some hidden mental phenomena, then such a relaxed attitude to interpretation is not acceptable. But it is the very starting point of the version of interpretivism propounded in this book that we should not make this assumption. I think that my condition is preferable to Dennett's predictivity criterion, since predictivity and the possession of mental states often come apart. First, not all ascriptions are made with the purpose of predicting behaviour and not all ascriptions need to generate predictions. The basic motive for mental state attribution is to make sense of the subject's behaviour and other mental states in mental terms. Second, not every behavioural prediction involves the ascription of mental states and contents. Behaviour can be predicted in terms of an organism's capacities, as well as normative regularities governing action in certain contexts. In the case of some regular types of behaviour, it can even be predicted inductively, as Gauker (2003a) has it. Hence, if reliable prediction need not involve mental attributions at all, then why should we assess the acceptability of interpretations in terms of prediction?

---

11. Dennett (1987: 52, n. 4) outlines two kinds of tests for folk-psychological predictions: "action predictions it tests directly by looking to see what the agent does; belief and desire predictions are tested indirectly by employing the predicted attributions in further predictions of eventual action."

I conclude this section with a clarification of the notion of ascribability. The theory is given a weaker formulation in terms of ascribability, compared to the stronger but implausible formulation in terms of actual ascription. For, surely, the examples of actual ascription in everyday life that satisfy the canonicality requirements are not that numerous. But when we use the 'ascribability' formulation, then we need to lay down the constraints on this notion. One way to spell out the '–bility' part of ascribability is to do so in terms of possible worlds. In general, to say that *As* is ascribable is to say that there is at least one possible world in which *As* is ascribed. However, we are interested in worlds that are close to the actual world, for it does not make sense to explicate the possession of mental states in the actual world in terms of what happens in some far-away worlds.[12] The range of possible worlds that we consider when we say that *As* is ascribable comprises worlds that are similar to the actual world in all respects relevant to the possession of mental phenomena with the only exception that whereas in the actual world, *As* was not ascribed, *As* was ascribed in those worlds. It might take a long time to list all aspects that are relevant, but necessary conditions are that the ascription sources are identical (or very similar) to sources in our world, that the ascribers do not have any superhuman mindreading powers or the like, and that there are no differences in folk psychology and our concepts.[13] Let us mark the range of possible worlds that satisfies these similarity constraints, C, as the set of w_c worlds, W_c. The claim of the ascription theory, put informally, is thus that one has mental states if and only if these states are canonically ascribed to one in the w_c worlds. To say that such and such mental states are canonically ascribed in the w_c worlds is to say that the ascription in those worlds satisfies the canonicality conditions, the coherence condition and the condition concerning revision. The last condition is that if the ascription were in fact made, it would not require revision. This can be translated into the present scheme as follows: there is a range of possible worlds W_c such that in every nearby world w_c belonging to W_c in which *As* is ascribed, it is not found to require revision. In this formulation, the modal locution 'would not require revision' transforms into the factual locution 'is not found to require revision (in the relevant range of worlds)', and this allows us to get rid of the second counterfactual.

---

12. Those philosophers who follow Lewis (1983b) in regarding 'actuality' as an indexical notion should perhaps replace 'actual world' with some non-indexical term denoting our world in the above formulations.

13. Another important constraint is that there are no interfering factors in the world where *As* is ascribed. This rules out factors that prevent revision of an ascription that requires revision, such as the whole world being destroyed once the ascription has been made or other obstacles invented by an evil demon (cf. Baker 2003: 187).

Finally, we are in a position to explicate canonical ascribability. To say that *As* is ascribable is to say that there is a range of possible worlds W_c such that *As* is ascribed in the w_c worlds. The ascription is canonical if and only if *As* is a statement in approximation to the maximum coherence with the ascription sources in the w_c worlds and in every nearby world w_c belonging to W_c in which *As* is ascribed, it is not found to require revision.

However, this account of ascribability is not completely helpful since it does not yet explicate what it is that makes *As* ascribable in that world where it is ascribed. Although 'ascribability' is a modal notion, it must be rooted in actual facts. From this perspective, there are two kinds of facts that make it the case that certain mental predicates are ascribable to one. First, there are facts about the subject and its environment – facts about how they actually are, as listed in the ascription sources. Since the ascriber is ineliminable from the theory, the second relevant class of facts consists of facts about canonical ascribers. How the canonical ascriber takes the subject to be is an ineliminable part of the interpretation, for it is the ascriber who applies mental vocabulary to that subject and places its action in the network of folk-psychological interconnections. It is true that the act of ascription itself need not be actual, but all that matters is that canonical ascribers can be actual, that is, ordinary interpreters can instantiate the (pleonastic) property of being a canonical ascriber. To explain the 'can' that is involved here we need not jump into the metaphysically deep waters of possibilities either, for given the relaxed standards of canonicality, all of those who apply folk psychology in the everyday interpretation of others fulfill the 'can' condition. The only differences are that in ordinary interpretation, there is usually less data available and perhaps less attention is paid to the coherence between the sources and the ascription.

In this way, ascribability is brought back to earth. A statement is ascribable in virtue of facts that are collected in ascription sources and facts about how ordinary ascribers interpret subjects.[14] Due to the uncodifiability of interpretation, a full explication of the latter facts – the capacities of ascribers – is most likely unfeasible. This is discussed in the next section.

14. It might seem that by making the ascriber's capacities part of the explanation of the possession conditions of mental properties, the account has become circular. I rebut this objection in the next chapter.

5.5 The indeterminacy and uncodifiability of the mental

In what follows, I clarify the stance of the ascription theory on the indeterminacy issue. Since indeterminacy is related to the controversial question of whether rationality is uncodifiable and may be a result of the latter, let us begin the discussion with the issue of the codifiability.

According to one reading of Davidson's anomalous monism, one reason (among others) for the denial of strict psychophysical laws is that rationality cannot be codified. To deny such laws is to claim that there are no strict rules for predicting and deducing individual's mental states from premises couched in physical terms. Some authors make the further point that "there is no definite set of rules or principles for arriving at the best interpretation of an agent" (Child 1994: 59). Note that this is a stronger statement than that required for Davidson's anomalous monism, which requires only that an agent's mental profile cannot be deduced from the set of *physical* truths. Nevertheless, some textual support for the stronger claim in Davidson's writings can indeed be found (for example, Child cites Davidson 1985b: 245 and 1986: 173). But I shall not concentrate here on the question of what Davidson did say or should have said, and proceed with the stronger reading of uncodifiability spelled out by Child:

> There is no fixed set of rules or principles from which, together with a statement of the circumstances of any particular case, we could deductively derive a complete, detailed specification of what one ought to do or think in that case.
> (Child 1994: 58–59)

Child (1994: 60, original italics omitted) points out that it follows from the uncodifiability of rationality that we have no "principles for deriving a statement of what attributions of mental properties make best sense of *S*", and in the interpretivist picture, this means that "there are no principles for deductively deriving a specification of *S*'s mental properties."

In what follows, I discuss two questions: first, what is the stance of the ascription theory on uncodifiability?; and, second, is uncodifiability a problem for interpretivism? I do not restrict the discussion only to rationality, and talk about the uncodifiability of interpretation or the uncodifiability of the mental in general. I agree that interpretation is uncodifiable in the sense that there is no single fixed set of principles suitable for the deduction of a subject's mental profile. The effect of uncodifiability may seem to be even stronger, given that the attribution of mental states is not guided by strong rationality assumptions. For even if one were to establish a set of rationality principles, they would be of no use in interpreting an irrational subject.

However, one may think that uncodifiability and interpretivism cannot be supported together, for if uncodifiability holds, then interpretation cannot be effected. And if interpretation cannot take place, the very idea of interpretivism seems to lack foundations. I do not share this pessimism. The uncodifiability thesis can be read as saying that there is no *explicit* and *complete* set of principles for deducing a subject's mental states. It does not say that there are no such principles or that we cannot be aware of such principles or that we cannot rely explicitly on some principles of interpretation. Neither does it say that once we have come to an ascription of a mental state to an agent, it is not possible to construct a story about our reasons for such an interpretation in folk-psychological terms. Uncodifiability does not rule out the option of having a certain implicit competence for applying folk psychology and thus coming to various ascriptions, some of which may be up to the standard of canonicality. The fact that there is no straightforward way to make explicit what principles this competence relies on is not damaging to interpretivism, unless the latter specially requires such lists.

Child (1994) does not attempt to give a straightforward proof for the uncodifiability of rationality, but he points out that there is similarly no codification of rules for practical rationality, for choosing between different theories, and for aesthetic judgements. In his view, the burden of proof is on those who assume that there is a codification of rationality and the principles of the related areas. However, the case against the full codification of the process of interpretation can be made more strongly by reference to certain features of mental discourse that show why the codification is not feasible. The uncodifiability claim says that there is no fixed set of rules for calculating a specification of a mental profile from some initially specified conditions. Chater and Pickering (2003: 348–350) mention two main features of folk psychology that preclude a full and explicit formalisation of our knowledge of it and turning this into a feasible computer algorithm, at least on the present understanding of computation. First, no subdomains in folk-psychological knowledge are safeguarded from further defeaters. Using a term from Fodor (1983: 105), they say that folk-psychological inference is *isotropic* in the sense that the conclusion is open to defeat by new information from a variety of sources. Second, they say that common-sense inferences are *fractal*: the complexity of every further inferential link is comparable to the complexity of the whole inference. In their view, these properties present difficulties for attempts to formalise folk-psychological competence. This may be one explanation as to why there is no full and complete codification of folk psychology and its application in the offing.

Note that the explication I have given to what is involved in ascription does not amount to a full codification. Although I tried to make everything as precise and explicit as I could, the force of uncodifiability considerations cannot be denied.

For example, although there are certain rules of thumb governing the sources, no fixed list of principles can be provided. Hence, there is no explicit and complete algorithm for proceeding from the description of the information contained in the source to the formulation of the ascription statement. Furthermore, there are no strict rules for assigning measures of relevance to the sources. It follows that there is no explicit formula for deducing the attribution of a mental profile from the set of initial conditions. However, this is not a defect of the account, given the grave circumstances. Certain conditions are made part of the account that balance uncodifiability. First and foremost, the ascriber is ineliminable from the account. Since no complete algorithm has been made explicit, the process of interpretation is dependent on the implicit capacity of the (canonical) ascriber. It is the ascriber that judges the sources of information and assigns them a measure of relevance. If this capacity is at some point made explicit – although, in light of the aforementioned considerations, I have strong doubts about the feasibility of this – then the ascriber will become redundant and be replaced with an enormously complex algorithm. The resulting account would still be constitutive interpretivism, since interpretability as a condition of possession would be still ineliminable. After all, it is an algorithm of interpretation that has been formulated. Mental properties would still be interpretation-dependent properties. The only difference is that the process of interpretation would then have been made completely explicit. The second condition that stops the possession of mental states being completely beyond our ken is the slack in the formulation of the canonicality clause. The stipulation about revision gives canonical status to a variety of ascriptions which pass the relaxed criteria. Hence, there is no need to worry that uncodifiability *cum* interpretivism leads to the conjecture that, since there is no codifiable way to fix our mental profiles and since our mental profiles depend on such fixing, they are either completely undetermined or we do not have mental states at all.

The uncodifiability of interpretation is related to the indeterminacy of the mental. The uncodifiable elements of interpretation lead to the variance in ascriptions and this is usually seen as a sign of the indeterminacy of the mental. Recall from the discussion in Chapter 3 that Davidson regarded indeterminacy as relatively harmless to his realism concerning the mental. His view is that indeterminacy exists only between the alternative theories of interpretation. The choice between these theories may be arbitrary, but this does not really matter, since different schemes are just alternative ways of stating the same facts. It follows from this that there are invariants between the interpretive theories, hence, there is no serious indeterminacy internal to a theory. If there were internal indeterminacy, that is, if one theory left room for different assignments of beliefs and meanings, there would not be any invariants. Dennett's view on indeterminacy is more radical, for he allows that acceptable interpretations can differ substantially, and thus

there need not be any invariants. However, he does not explain what the sources of such indeterminacy are.

On the one hand, the ascription theory has resources to provide such an explanation, but, on the other hand, the indeterminacy issue itself is transformed. It may seem that indeterminacy cannot be denied once uncodifiability is recognised. However, the issue is more complicated. In the model outlined in previous sections, there are two main points of variation: the declarations of the sources and the relevance of the sources. A given source can be assigned different declarations and it can be assigned different relevance measures. Given such alternative assignments, the resulting ascriptions can come out differently as well. An additional source of variation is the coherence measure. Perhaps it is common to presume that there can be no variation in the coherence measure. Although there may be different models for calculating coherence, there is a strong intuition that once the elements are fixed, there is one best way in which these elements fit together. All the ensuing differences are thus artefacts of the models and not a feature of the situation. However, since, in my account, those ascriptions that approximate to the maximum coherence can be canonical as well, this may lead to differences in ascriptions and it is thus an additional source of variance.

It is a feature of my account that as long as different ascriptions pass the standard of canonicality, the subject has the ascribed mental states. The possession of such states is as determinate as it can be. There is no uncertainty concerning whether one has either the state s_1 or the other state, s_2. If they can both be canonically ascribed, the subject has both states. Does this mean that I deny the indeterminacy of the mental? We need to distinguish between the variance of interpretations and the indeterminacy about the possession of mental properties. On the one hand, interpretations of one's mental profile can vary, provided there are different readings of the ascription sources and their relevance as well as the possible fluctuations in the coherence estimation. On the other hand, the notion of canonical ascription helps us to get rid of the indeterminacy at the level of possession of mental states. When it is allowed that all variant canonically ascribable mental states are possessed, then the variation in interpretations does not entail the indeterminacy of the mental. In prevailing discussions, it is claimed that we have a case of indeterminacy when we have competing ascriptions but there is no evidence to decide between them, even "when all the evidence is in" (Davidson 1973a: 327). In this respect, what matters is not the mere fact that we do not have access to a sufficient amount of evidence for interpreting a subject, but rather the fact that all the possible evidence is not sufficient for determining what mental states one has. Since the discussion takes place on the interpretivist background, this transforms into a claim about indeterminacy concerning the very possession of mental properties. Still, the discussion has presumed that the subject cannot

have all the alternative mental states or profiles, that we somehow have to choose which one the subject has, and, if we are unable to choose, then we have to be content with indeterminacy. However, once it is allowed that one has all alternative states, indeterminacy evaporates. Note also that the idea that a large number of mental states is canonically ascribable to one is not counterintuitive, once we have dropped the requirement that for each distinct mental state there should be a corresponding neural configuration in the brain. (Cf. Dennett's (1988c: 537) suggestion that there is an infinite number of beliefs in a finite brain).

At this juncture, one may develop the following objection. Imagine that we can ascribe very different mental profiles to an agent. The ascription theory is then committed to the claim that the agent has distinct mental profiles. This runs contrary to common intuitions, for it could make the total set of the agent's mental states absolutely incoherent. Perhaps the most obvious answer to this objection is that ascriptions are always made relative to the interpretive scheme or manual and when we estimate the consistency of an agent, we should not mix manuals (cf. Føllesdal 1990: 105). However, given uncodifiability considerations, it is hard to individuate interpretation manuals. For example, if we change the estimation of the relevance of the information, do we have a different manual? If we change the declaration of a source, are we interpreting in accordance with a different manual? The idea of a manual presumes that interpretation is fairly automatic, so that once we have made the choice between different manuals, the deduction of the ascription follows mechanically. Uncodifiability considerations tell us that the situation is not that straightforward.[15]

In my view, a better answer to the objection is to point out that there are reasons to expect that *incompatible* mental profiles are not canonically ascribable. Although the canonical ascriber can assign different values to the aforementioned variation points, he is constrained by the existing information. The consideration of the sources is not arbitrary and neither is the estimation of their relevance. The sources are not considered in isolation, but on the backdrop of other sources, and the declarations of the sources can be respectively revised in view of overall coherence. Hence, it is difficult to imagine the possibility that the ensuing ascriptions would turn out to be incompatible.[16] Another important constraint is

---

15. This tension between uncodifiability and the notion of the interpretive scheme makes it difficult to grasp how Davidson can liken interpretive schemes to systems of measurement given that he countenances uncodifiability.

16. Note that the talk of "incompatible ascriptions" comes from Davidson's (1997: 120) consideration of Dennett. Dennett (1991b: 49) himself talks about systems that yield substantially different ascriptions without clarifying what a "substantial difference" is and without giving any real-life examples.

that the ascription is canonical only if there is no need to revise it. The fact that the canonical ascriber is faced with two different or incompatible ascriptions is a good reason for seeking further revisions. In that case, substantially different ascriptions are not both canonical. Given these considerations about the constraints of the interpretive process, my contention is that incompatible ascriptions do not present an obstacle for the ascription theory.

This concludes the presentation of the ascription theory. It will be further qualified in later chapters, especially in Chapters 8 and 9, where I apply it to perceptual states and self-knowledge. The ascription theory offers an improvement on existing interpretivist accounts by providing a more detailed account of what is involved in interpretation. In the next chapter, we will see how it fares in respect of common objections to interpretivism.

CHAPTER 6

Objections and defence

There is a standard battery of objections directed at various forms of interpretivism. The first part of this chapter is devoted to the defence of the ascription theory against some such common objections. As similar objections can be presented to various forms of interpretivism, they do not constitute criticism of the specific details of the ascription theory. Still, some of my answers depend on the details of the theory. One may take the fact that the ascription theory has resources for rebutting various objections as an indirect reason for its support. While the chapter begins with a rebuttal of objections, the second part investigates the options for developing an argument in support of interpretivism. The chapter ends with a formulation of one such argument.

6.1 Objections

As noted, the objections discussed in this section can be (and often are) directed at interpretivism in general, not only at the specific version propounded here. However, I formulate the objections in a manner pertinent to the ascription theory, and to keep the discussion at a reasonable length, I generally do not go into the issue of how Davidson or Dennett might respond. I shall consider five objections concerning circularity, regress, liberalism, intrinsic intentionality and observer-dependence and answer them from the point of view of the ascription theory.

6.1.1 The circularity charge

The most common objection to interpretivism is the charge that it is circular. Since interpretivism explains the possession of mental states and contents in terms of ascriptions made by an interpreter, this explanation presumes the notion of mental states – the states of the interpreter. Put otherwise, the charge is that ascribing, stance-taking and interpreting are intentional states themselves and, therefore, they fail to provide an illuminative account of intentional states. Or in another version: if the interpretivist aims to provide a fictional or instrumental account of all mental states, then he cannot give the same reading to his own acts

of interpretation. This objection is presented or discussed in various versions by Brandom (1994: 58–62), Child (1994: 53–54), Hornsby (1997: 179–180), Seager (2000: 122) and Wright (2002: 10).

Note that the ascription theory is not committed to what can be called 'definitional circularity', for it does not claim that interpretivist elements are involved in the definitions of mental terms or that they play any part in explaining the meanings of mental terms. Hence, the ascription theory is not vulnerable to the charge that the explication of the notion of belief involves interpretation and that it is not possible to explicate the notion of interpretation without the notion of belief. However, the circularity charge can be pressed against the ascription theory at the level of possession: the explication of what it is to have a belief involves mention of interpretability, but interpreting would in turn involve having beliefs.

My reply is that the circularity charge is serious when we assume the widespread attitude towards the mental. This attitude presumes that there are mental properties, including intentional properties, and that the only way to uphold physicalism is to naturalise mental properties by reducing them to physical properties or giving a constitutive account of mental properties in non-mental terms. I do not share this attitude. As I already explained in Chapter 2, I do not assume that there are mental inflationary properties and I do not assume that mental properties (or states) present ontological problems. All we need to clarify is the role of mental concepts. In view of the dispiriting track record of the project of conceptual reduction, I do not think that a satisfactory reduction of mental concepts to non-mental concepts is forthcoming. The best we can attempt is to explicate intentional notions and clarify their application conditions in terms of other intentional notions that are more familiar and less puzzling. However, as noted, the prospects of this enterprise are much less gloomy than those of the ontological task of dealing with irreducible mental entities. Since the case for mental *entities* in the inflationary sense has not been made, there is no reason to succumb to this, more serious concern.

A related aspect of the idea that leads to the circularity charge is the assumption that intentionality is a real "power" in the world that can make things happen. This is the idea that there exists "real intentionality" that stands in need of explanation. I discuss this idea in Section 6.1.3. Here it has to be noted that when the interpreter is ascribing mental states to an agent, he is not exerting some mysterious power that remains unexplained or which cannot in turn be explained without exerting the very same power. There is no such thing as intentional power. All the causal powers of the interpreter are physical and when the interpreter is making an ascription, he is exerting the physical powers of

his brain states. The power that makes things happen in our physical world is itself something physical. The behaviour of interpreters has a physical basis and, therefore, when one performs the act of ascription, the act is not one of the *sui generis* kind, or one caused by the unexplained power of intentionality. It is effected by the underlying neural processes. This is not to reduce the acts of ascription to the ascriber's brain processes, however. If one were to do this, the adjacent interpretivism would become redundant, for one could then presumably reduce the mental states of the interpreted subjects to their brain processes as well. The idea is rather that as a vocabulary, intentional notions cannot be explicated in non-intentional terms and can only be explained at the intentional level, as is, for example, done in interpretivism. But this does not preclude us from giving a causal story of the underlying natural processes that take place in us and our interpreters, and which are those efficacious processes that make things happen. For more on this theme, see Chapter 7.

6.1.2 The regress objection

The regress objection is related to the circularity charge and sometimes they are not even properly distinguished. It is the charge that the ascription, being a kind of mental state, must in turn be ascribed if one wants to uphold the claim that all mental states are ascription-dependent. This objection claims that ascription could never get started, for it leads to a regress of ascriptions. For example, Georges Rey presents this objection to Dennett's approach, saying that if intentional predicates are specified relative to an interpretation, then they cannot even be specified, for the specification of '*x* believes *p*' has to look like the following: "*x* believes *p* relative to intentional system *y* relative to intentional system *z* relative to …" (Rey 1994, 282; see also Rey 1997: 79–80). This objection is directed against Davidson's interpretivism by Baldwin (1993: 197–198). It is also forcibly voiced by John Heil:

> If my thought's having a certain content requires that it be described, interpreted, theorised about in a certain way, how are we to understand *these* activities except intentionally. A describer, theoriser, or interpreter must think about, designate, or refer to what he describes, theorises about, or interprets. Such activities either require some further act of interpretation or they do not. If they do, then a regress looms. If they do not, then we are owed an explanation of how are we to take them. The regress can be blocked, perhaps, by opting for a broader antirealism, one that encompasses, not merely the intentional attitudes, but the nonintentional world as well. …antirealism might be thought to represent a conceptual last resort.
>
> (Heil 1992: 229–230)

I do not think that one has to adopt a global antirealism in order to stop the regress, nor do I see how antirealism can help in this regard. My answer does not presume antirealism about the "nonintentional world". As with the previous objection, the problem of specifying mental predicates does not concern the ascription theory; the regress threatens rather at the level of possession. Namely, it seems that the very having of a belief sets off an infinite regress of ascriptions:

> *x* believes that *p*
> *y* ascribes (the belief that *p*) to *x*
> *z* ascribes the state of ascribing (the belief that *p* to *x*) to *y*
> *a* ascribes the state of ascribing (the state of ascribing (the belief that *p* to *x*) to *y*) to *z*
> and so on…

In the end, it is hard to see how anyone can have beliefs at all.

I do not think that this is a problem for the ascription theory. First of all, my account does not require actual ascribing, so the worry that possession cannot get off the ground is obliterated. Whereas ascribing beliefs is a state that in turn requires ascription, the fact that a belief is ascribable to one is not a mental state. When possession is connected to ascribability, it can easily be seen that there are objective facts about ascribability. The fact that belief that *p* is ascribable to one is a fact that holds independently of any actual ascription. As long as the subject satisfies the conditions for the ascription of the belief that *p*, this state is ascribable to the subject and there is no vicious regress that precludes one having the state. Indeed, it is true that for each mental state there is an infinite list of metafacts about the ascribability. But the same is true in other areas as well and it is not regarded as problematic. For example, if I think, it is presumably thinkable by me or someone else that I think and it is thinkable that it is thinkable that I think. Such metafacts about thinking can be constructed without end, and they do not cast any doubt on thinking.

It may be objected that since the ascription theory makes having a belief dependent on ascribability, these two cases are not analogous. In the case of the ascription theory, any ascribability is constitutively dependent on further ascribability. However, the dependence that is in play here is only conceptual; it does not bring any further inflationary entities into existence. By the same token, it cannot fail to bring any inflationary entities into existence either. Facts about the possession of mental states are pleonastic facts, which hold independently of what we may actually do. Although the explanation of their holding involves reference to ascribability, this does not prevent them from holding. Alternatively, it may be asked that if we explicate ascribability in terms of there being an ascriber in some possible world, then does not the regress arise again with respect to this possible ascriber? In virtue of what does she have her mental states? I have some

comments on this. First, the explication of ascribability in terms of other possible worlds does not provide the ultimate illumination; it only paraphrases the notion of ascribability without providing any substantial explanation. What makes it the case that a certain mental state is ascribable are facts about the actual world. Second, even if we accept this analysis, then there is no need to be realists about possible worlds. In the latter case, the issue of the interpretation of a possible interpreter cannot give rise to the worry that the possession of mental states involves any real regress.

6.1.3 Intrinsic intentionality

A further objection to interpretivism is that it does not explain intrinsic intentionality. The idea is that artefacts like sentences in books, signals, tokens in computer code mean what they do in virtue of the meanings we have given to them. It is said that the intentionality of artefacts is derived intentionality and comes from the intrinsic intentionality of people's mental states (e.g., Searle 1992). I presume that Dennett (1987, 1990a, 1990b) has done enough to show that the very idea of intrinsic intentionality is a myth. It is one thing to maintain the distinction between people, books and computers, and it is another thing to invest in the obscure notion of intrinsic intentionality that somehow originates from us and is not derived from anything else. Dennett has used the example of a memorised shopping list as a compelling intuition pump: "A shopping list written down on a piece of paper has only the derived intentionality it gets from the intentions of the agent who made it. Well, so does a shopping list held by the same agent in memory!" (Dennett 1996: 52). Elsewhere he elaborates on this example:

> A shopping list in the head has no more intrinsic intentionality than a shopping list on a piece of paper. What the items mean (if anything) is fixed by the role they play in the larger scheme of purposes. We may call our own intentionality real, but we must recognize that it is derived from the intentionality of natural selection, which is just as real – but just less easily discerned because of the vast difference in time scale and size. (Dennett 1987: 318)

Dennett claims that our intentionality is derived from Mother Nature or natural selection, but Mother Nature has intentionality only metaphorically, viewed through the lenses of the intentional stance. Eventually, his point is well conveyed by his paradoxical-sounding claim that "derived intentionality can be derived from derived intentionality." (Dennett 1996: 54). But this makes sense on the background of his deflationary approach to intentionality.

Sometimes the assumed task of explaining intrinsic intentionality leads to the search for a naturalising account of mental representations. The latter are

supposedly the origin of our meaning-conferring powers. However, it is important to note that when interpretivism rejects the idea that the intentionality of our beliefs and desires derives from the intentionality of our mental representations, it does not have to reject the very notion of an internal mental representation. There are explanations in cognitive science that routinely make use of internal representations and subpersonal information processing. These can be accommodated by the ascription theory as long as their possession is taken to be ascription-dependent.[1] Their ascription could be based on some psychological theory and the ascription sources would be respectively different from those of common-sense ascription. Apart from such ascription-dependent intentionality, no mental "representation" is intrinsically a representation (see also Elton 2003: 164–165). Neural systems are not intrinsically representational; they can detect changes or properties relevant to their structural constitution, but the detecting differs from representing. The detector detects, but does not encode the stimulus. This idea is well expressed by Dan Lloyd (2000: 191): "From the detector's point of view, something happens, and the detector is blind to any 'meaning' or 'reference' in its state change beyond the brute fact of being in a new state." If we can explain all that we need in terms of ascription-dependent mental representations and property-detectors in the brain, the notion of an intrinsically intentional mental representation is left without impetus. Of course, it is a separate task to show that this is indeed the case and this cannot be undertaken in the present book. To conclude on a positive note, the upshot is that interpretivism presents a compelling alternative to the assumption that intrinsic intentionality is required for cognitive-scientific explanations.

6.1.4 Liberalism and chauvinism

Another common objection to interpretivism is essentially the charge that interpretivist attributions do not line up with our common intuitions concerning mental state possession. There are two sorts of charges of this nature. On the one hand, there is the accusation that intepretivist attributions are too liberal, for interpretivists are committed to ascribing minds to things and creatures that do not have minds according to folk intuitions. On the other hand, there is the opposite

1. Cf. Dennett's (1982: 354) somewhat neglected remark: "The representations posited by the new cognitivism are not deemed to be the ultimate and underived and incorrigible font of all meaning or intentionality. … Just as the semantical or hermeneutical interpretation of a scrap of text found incised on a wall or scribbled on a page depends on attributions of intentionality – belief and desire and so forth – to users and makers of the text so the semantical interpretation of a bit of 'mental representation' in the brain – a found inscription in the language of thought – depends on exactly the same family of intentional attributions."

charge of chauvinism: interpretivists cannot ascribe mental states to creatures that most probably do have mental states.

One version of the liberalism charge is Peacocke's (1983: 205–206) Martian marionette thought experiment. We are invited to imagine a human body with no brain. In virtue of being remotely controlled by Martians, it behaves just like any normal human being. This is taken to be a problem for Dennett's intentional system theory, for it fulfils the condition of behavioural prediction. Its behaviour is reliably predictable and hence the Martian marionette is a subject of intentional ascription. However, the folk intuition is that such a radio-controlled marionette has no mental states whatsoever. This objection is not insuperable for an interpretivist, for one may set the requirement that behavioural dispositions need to have some internal causal basis without compromising the interpretivism (McLaughlin and O'Leary-Hawthorne 1994: 215). One can thus answer that the ascription of intentional states to the Martian marionette has no explanatory power, since we cannot explain the marionette's behaviour as an outcome of its having certain beliefs and desires.

However, such an answer leads to a further counterexample, which is known as 'Blockhead' (see Block 1981). This objection strives to establish the point that not only do we need to allow that behaviour has some internal basis, the *structure* of this internal basis is also crucial. Blockhead is a creature that exhibits the right kind of behavioural responses to certain inputs, but instead of having a normal brain, it operates on the basis of a reference table that contains output routines for all possible stimuli. In this case, there is observable behaviour and the internal basis for this behaviour, and still there is a strong intuition that Blockhead is not a thinker and cannot be ascribed mental states.

Dennett (1993b, 1994) has sometimes dismissed the Blockhead case as nomologically impossible, refusing thus to follow the widespread method of studying concepts by imagining conceptually or metaphysically possible scenarios. On other occasions, however, he has conceded that granting the possibility of Blockheads, they would indeed have mental states, since the data stored in the reference table instantiates the required internal structure in virtual form (see Dennett 1993b: 924). His answer is thus that although Blockhead's internal architecture is not physically and functionally of the right sort to act as a basis for behavioural dispositions, this architecture still encodes a certain complex structure. Elsewhere, Dennett reiterates his commitment to internal structure:

> a brain filled with sawdust or jello couldn't sustain beliefs. There has to be structure; there have to be elements of plasticity that can go into different states and thereby secure one revision or another of the contents of the agent's beliefs.
> (Dennett 1998: 329)

It might look as if this is a step away from his interpretivism, which regards reliable prediction as constitutive for the possession of mental states, but this is not the case. Right after his striking claim that "*all there is* to really and truly believing that *p*… is being an intentional system for which *p* occurs as belief in the best (most predictive) interpretation", he goes on to say that "this apparently shallow and instrumentalistic criterion of belief puts a severe constraint on the internal constitution of a genuine believer and thus yields a robust version of belief after all" (Dennett 1987: 29). Indeed, an interpretivist does not have to assume that the subjects of our mentalistic ascriptions have no complex internal structures. On the contrary, any system that exhibits behaviour complex enough to be explained in intentional terms has to have a complex internal structure. The point is rather that such internal structure does not determine what states are ascribable to one and therefore does not fix the states one has.

However, most philosophers would refuse to ascribe mental states to Blockheads and it can be argued that Dennett does not really have to bite the bullet and allow thought to Blockhead. For example, Elton (2003: 197) proposes that besides the behavioural criteria, there are also proscriptive constraints on belief ascription. There is the "no miracles constraint", which precludes all systems with miraculous internal processes, and the requirement that the internal structure should be of a reasonable size. In Elton's proposal, such constraints would rule out Blockheads and comparable systems as candidates for belief ascription.

I think that there is room for improvement on these answers. On the one hand, it is hard not to notice the *ad hoc* nature of Elton's proscriptive conditions. Besides, it is not clear if there is any non-question-begging account that settles which size is "reasonable" for an intentional system. On the other hand, Elton's constraints point towards a more general solution. As mental state ascription is guided by folk psychology, certain proscriptive constraints can indeed be part of the common-sense ascription conditions. They influence the eventual outcome of interpretation through the background source. We can see that the Martian marionette and Blockhead present problems only to versions of interpretivism that proceed from an austere evidential basis. If we allow that besides behaviour, ascription can rely on various sources of evidence, the problem of excessive liberalism disappears. Once we allow that facts about a system's background (e.g, that its body is controlled remotely or that its behaviour is based on a reference table) can also guide interpretation, the theory is not committed to ascribing mental states to subjects that do not comply with folk standards.[2] In this way, the common-sense

2. One of Child's (1994: 45) responses to the Blockhead thought-experiment relies on the causal history of the system. According to Child, we cannot ascribe thought contents to Blockhead, for it lacks previous causal interactions with the environment on which to ground

disinclination to ascribe mental states to Blockhead and Martian marionettes would be accommodated in the theory. Similar considerations provide an answer to the question concerning the grounds we have for ascribing mental states only to certain subjects and not to others. Dennett's criterion of predictive success, although robust, proves to be exceedingly liberal. I would prefer to answer that since mental predicates are ascribed from the folk-psychological standpoint, this perspective has already circumscribed the range of potential subjects of interpretation. There is no deeper or substantial explanation to this fact. It is part of folk psychology that people have beliefs whereas tables and lecterns do not. It is not so clear whether animals have beliefs according to our folk representation. This can be discovered if we study our habits of ascription. One can try make explicit certain conditions that the subjects of ascriptions have to fulfil, but the prime source for such conditions should be our folk psychology. There are cases where folk psychology fails to provide a clear-cut answer. Although Blockheads are ruled out from the range of subjects of intentional ascription, matters are less clear with more complicated robots, presently encountered only in science-fiction movies and books.[3] Presumably, such stories stretch and shape our notion of a person and if the time comes for a real-life encounter with such artificial creatures, then perhaps folk intuitions will have become determinate enough to settle the question of mental state ascription. What is crucial is that on the interpretivist approach, such issues are settled by consulting and developing our folk-psychological representations; they are not settled by discovering deeper facts about subjects. Although we have beliefs and Blockhead does not, there is no hidden ingredient that makes us real believers and that is lacking in Blockhead's head.

The opposite charge to liberalism is that of chauvinism.[4] While the previous thought experiments held outside behaviour fixed and varied the internal structure, let us consider now the scenario that holds the internal states fixed and removes behaviour. Yet again, the critics' aim is to show that a theory of

---

his dispositions to think about the objects in its surroundings. This reply, as well as Davidson's (1987a: 443–444) similar approach to Swampman, presume a teleological account of content. Although interpretivism can in principle be coupled with teleological considerations, one needs to explain how such coupling is possible without undercutting the very point of interpretivism. If causal-historical factors determine content, then why need interpretation at all?

3. What would folk say about the main characters in popular books like *The Positronic Man*, by Isaac Asimov and Robert Silverberg (1992), or films like *Artificial Intelligence, AI* (2001), directed by Steven Spielberg, and *I, Robot* (2004), directed by Alex Proyas?

4. This should not be confused with the charge of the same name, often directed at type-identity theories: since they identify mental state types with types of brain processes, they are chauvinists in respect of the material of which believers and thinkers can be composed.

mental state possession must take internal states into account. We are invited to consider the case of someone with complete paralysis, having no capacity for peripheral movements.[5] Can we ascribe mental states to such a person and on what grounds? Intuition supports the verdict that a paralysed person has mental states, lacking only the ability to express them. If the ascription theory is committed to the claim that the person has no mental states, then something must be wrong with the theory.

Let us first see what Dennett has to say about paralytics:

> What about real cases of peripheral paralysis? First, the only real cases have to be people who have lived an unparalyzed life for years – all other imaginable cases are cow-sharks, only logically possible and rudely dismissable. Second, the persistent integrity of the internal structures on which their continuing mental lives putatively depends is not a foregone conclusion. To the extent that the paralysis is truly just peripheral (unaccompanied by the atrophy of the internal), then, of course, such a sorry subject could go on living a mental life… But all good things come to an end, and in the absence of normal amounts of 'peripheral narrow behavior', mental life will surely soon fade away, leaving only historical traces of the vigorous aboutness its activities once exhibited. How long would it take? A gruesome empirical question, whose answer has no metaphysical significance. (Dennett 1994: 564)

Dennett thus concedes that paralytics have mental states, but notes that the mental life would probably fade away due to the condition of behavioural deprivation. This does not explain how paralytics could have mental states in the first place, since their behaviour is not predictable from the intentional stance. Neither can he extrapolate from the normal condition, saying that the paralytic has those beliefs and desires he would have if he were not paralysed, for surely being unparalysed would change one's mental states (cf. McLaughlin and O'Leary-Hawthorne 1994: 237).

I surmise that the ascription theory need not attempt to expand on folk psychology on this issue. And I do not think that folk have clear-cut intuitions here either. We do not really grasp the case of paralytics, but these cases seem so

5. McLaughlin and O'Leary-Hawthorne (1994: 240ff) introduce Manthra, an alien lifeform that behaves and has the appearance of a praying mantis. Manthra has two brains. The simple one controls its mantis-shaped body, while the other resembles a human brain but cannot control the movements of Manthra's body. The problem the Manthra case poses for interpretivism is that presumably it has an intelligent mind (due to its human-like brain) but there are no observable criteria for ascription. My reply to Manthra scenario does not differ from my judgement on the paralytics' case.

intuitively compelling because one can easily imagine oneself lying down, exhibiting no behaviour and thinking, feeling, imagining and entertaining beliefs. However, when confronted with actual cases of paralysis with absolutely no peripheral movements, folk psychology breaks down. This is to be expected, for our mental concepts have been developed to apply to normal everyday cases. In view of this, I am not sure that it can be claimed without begging the very issue that paralytics have mental states, for it is not clear whether they have mental states according to folk intuitions. On the one hand, it seems easy to imagine oneself lying still and thinking; on the other hand, this is not quite the same as imagining oneself being completely paralysed. If there is no way to fix content to a paralytic's mental states, there is no ground for maintaining that the completely paralysed person really has mental states. As with the case of advanced robots, this issue is left indeterminate by folk psychology.

However, the ascription theory has resources to make at least some sense of the unfortunate situation of paralytics. There are two types of ascriptions paralytics could make. If their brain structures are left intact as assumed (apart from those parts responsible for the paralysis), then they would surely have brain processes and some such brain processes would have phenomenal what-it-is-likeness. If a purely causal application of mental predicates also counted as a kind of application, then those paralytics who possess conceptual abilities could apply mental terms to themselves. Then it can be said that paralytics make self-ascriptions or at least have the potential for this. Such ascriptions would be triggered causally by underlying brain processes. However, the paralytic's self-ascription could still be canonical, especially in view of the fact that in the case of paralysis, no canonical ascriptions can be made from the third person point of view. Recall that the ascription is canonical if it coheres with the sources and does not require revision. Assume that for such a self-ascription, no declarations of the sources are considered (although they could still act as defeaters: the ascription could be different if some source became relevant). But as there are no declarations, there is no coherence problem to be solved and any ascription that fulfils the second condition – not requiring revision – would be canonical. Of course, this would be an extreme class of canonical ascriptions, but if the causal application of mental predicates is permitted, then we can accommodate such cases.

Another type of self-ascription that a paralytic could make is of a more reflective kind. If their sensory systems are intact as well, then a paralysed person could make a self-ascripion that takes into account the stimuli that affect her, other ascribable mental states and any background information that might matter. Only the behavioural source would be invariant as long as the paralysis lasts. Such a self-ascription could also be canonical, depending on the occasion, given that it is coherent with the sources and does not need revising.

6.1.5 Observer-dependence

A final objection that I will discuss is the worry related to observer-dependence (cf. e.g., Braddon-Mitchell and Jackson 1996: 152; Gauker 1994: 300ff). This is vividly expressed by Fodor and Lepore:

> The problem is that, if the Projectivist account of the (putative) interpretive element in belief attribution is right, then what *you* can believe depends on what *your interpreter* can say. But if anything is metaphysically independent of anything, surely your repertoire of potential beliefs is independent of anybody else's repertoire of potential speech acts. (Fodor and Lepore 1993: 71)

The projectivist theory at which this objection was directed is Stich's (1983) account that individuates mental content in terms of what the interpreter believes. Stich (1983: 88) argues that 'S believes that *p*' is roughly equivalent to the utterance of *p* and "S is in a belief state similar to the one which would play the typical causal role if my utterance of that had had a typical causal history". Indeed, such an account makes one's belief dependent on one's interpreter's conceptual resources for uttering that *p*, and this is a good reason for rejecting it.

However, this does not show that a comparable objection cannot be directed at the ascription theory. If what is found objectionable is the general idea that one's mental profile depends on interpretation, then most versions of interpretivism can be faulted for this. Note that an objection with a similar structure can be aimed at externalism as well: "if anything is metaphysically independent of anything, surely the content of one's beliefs is independent of one's surroundings." I think that the proper answer from the externalist point of view is to reject the assumption of independence and to provide reasons as to why we have thought in other ways. An answer from the ascriptionist point of view would similarly involve rejecting the independence assumption, but it would also draw some distinctions that aim to soften the contraintuitive consequences of this move.

The assumption that has to be rejected is that one's mental states and contents depend only on oneself. This assumption fails to take into account the dependence of one's states on one's surroundings. As interpretivism suggests, the way others come to appreciate one's mental states also plays a role in fixing what states one has. It does not follow, however, that the matter of mental state possession thereby becomes relative to the conceptual resources, intellectual level or interests of the interpreters. Since the possession is fixed by the canonical ascription, the varying standards of different interpreters do not have an effect on possession, if they fail to match up to the level of canonical ascribers.

It is important to note that although describability in mental terms depends on interpretation, the processes that take place in one's brain are not ascription-dependent. Making sense of them requires interpretation, of course, but the fact

that they occur is metaphysically independent of the observers. Perhaps this fact makes the ascription-dependence of mental properties easier to digest. By way of the independence of the neural events that occur in us, it provides a way to keep independence intuitions, at least in part.

These are my replies to some common objections that can be directed at interpretivism. Certain worries and doubts can also be raised in relation to such themes as mental causation and self-knowledge. The applicability of the ascription theory to these areas is discussed in the coming chapters. The treatment of all other objections remains for future discussion.

6.2 Arguing for interpretivism

After responding to the various arguments *against* interpretivism, it makes sense to inquire whether there are any good arguments *for* the view. As is often the case with broad approaches, one might simply assume its central claims and then go on to develop the view, outlining its consequences and rebutting objections, without considering whether there might be some more direct ways of establishing the position. This way of proceeding follows the example of Daniel Dennett, who says that: “I propose to yield a little ground and see where we are led” (Dennett 1987: 129). Of course, if one can show that the alternatives to interpretivism are prone to objections and that interpretivism can be elaborated and extended in various interesting ways, then this would itself constitute a positive, though indirect, case for the approach. This is also what I have attempted in this book. However, in this section, I attempt to investigate some ways one could argue directly for interpretivism. The ensuing discussion repeats some of the points made in previous chapters and it thereby also partly fulfils the summarising task.

What interests me is whether there is any general argument that could motivate the central interpretivist thesis, which makes the possession of mental states dependent on their ascribability – in accordance with certain nontrivial criteria for ascribabilty, of course, but we can omit this qualification for now. The argument that I seek does not depend on the more specific details of the ascription theory; I proceed rather from the general standpoint of intermediate interpretivism. In addition, in terms of the distinction introduced in Chapter 3, the interpretivism I am trying to support is a view on the possession of mental states and not necessarily an account of their nature. In a sense, if there were such direct arguments in support of the view, this fact itself would be quite striking. First, the interpretivist thesis is far from being uncontroversial. Second, it is not common in philosophy for there to be knock-down arguments in support of philosophical positions. Moreover, there are hardly any convincing cases in philosophy. Hence, I do not expect to find a fully-worked out deductive argument that will turn all readers into interpretivists.

By making explicit the lines of thought that can provide materials to argue for interpretivism, there are two pitfalls that are better avoided. One should steer clear of verificationism and behaviourism. Hence, if an argument includes some verificationist or behaviourist premise, it is unacceptable. There is a recurring suspicion that interpretivism incorporates some verificationist premise and, eventually, falls back on the forlorn behaviourist position. Hence, if interpretivism can be supported by arguments that are free from verificationist and behaviourist elements, or includes only some unproblematic versions of them, then we would ease such worries.

It is interesting to note that there is no place in the main works of Davidson and Dennett where a direct argument for interpretivism is presented. Davidson is fond of general, even transcendental, arguments, but his overall project is not to argue for interpretivism: it is rather to formulate a unified theory of meaning and belief. Dennett, on the other hand, prefers devising examples that change our intuitions to philosophising by making fine distinctions and presenting arguments and counter-arguments. In what follows, four considerations for the general interpretivist conclusion are examined. Some of these considerations, construed as arguments for the necessity of interpretability, are discussed by Child (1994: 32–39). I reject some of them and use some for the eventual formulation of the argument.

6.2.1 The consideration from behavioural criteria

It is commonplace to locate interpretivism in opposition to the so-called 'Cartesian picture', according to which mind contains private entities, ideas, which in addition to being metaphysically unique (or at least different from physical entities) also have special epistemic properties: they are given only to the person who entertains them, and the person has special authority concerning their mental states. Interpretivism is part of the current opposed to this approach, getting support from the epistemic claim that there are no such private objects. In some cases, such an anticartesian attitude can be backed up with the Wittgensteinian presumption that any ascription of unobservable properties should be based on something observable.

How do get from this attitude to interpretivism? Child (1994: 36–37) elaborates the idea in this way: it must be a feature of mental predicates that they can also be ascribed to other people. If ascribability to others were not essential to mental terms, then the primary use of mental terms would be private, in application to one's own experiences. These terms would then be applied to other people on the basis of analogy with one's own case. The problem with this strategy is that

there would still be a need for some behavioural criteria by which one could apply mental terms to others or even to single out the range of subjects to which mental terms are applicable (Child 1994: 37). The solution is then to take the application conditions of mental terms to contain reference to behavioural markers. Of course, the connection between behaviour and the applicability of mental terms need not be a necessary connection, but the behavioural constraints would still play a sufficient role. So this line of thought is not committed to the strong form of behaviourism, which postulates that the connection between mental states and behaviour is noncontingent, but it smacks of a certain weaker form of behaviourism nonetheless.

However, what we are looking for is an argument to show that the possession of mental states is constitutively bound up with interpretability. What we get from the considerations from behavioural criteria are the following claims:

1. Mental terms are not applicable solely to one's own case (the anticartesian line)
2. The application of mental terms is (partially) answerable to behavioural criteria (weak behaviourism)

This does not tell us anything about the nature or the possession of mental states. What we need minimally is an additional link between the public applicability of mental terms and the public constitution of mental states. In addition, we need to link constitution with interpretability. As it stands, these two claims are consistent with views that take facts about the possession of mental states to be completely ascription-independent facts. For example, these claims are consistent with functionalism that includes behavioural roles among the individuation conditions of mental states.

6.2.2 The consideration from the publicity of meaning

Let us reflect on the consideration that connects the publicity of meaning with the publicity of mental states. Child (1994: 32) attributes this to Davidson. The idea is that intentional states (especially beliefs) inherit their publicity from the publicity of meaning by being inseparable from the latter in the context of interpretation. As we saw earlier, Davidson's view is that beliefs and meanings must always be ascribed together, for we cannot determine a subject's beliefs, given the sentences he assents to, if we do not know which meanings the subject attaches to his words.

The claim that terms must be publicly applicable and the claim that meaning is public are different claims. Why assume the latter? Davidson (1990b: 314) clearly assumes that meaning is public: "That meanings are decipherable is not a

matter of luck; public availability is a constitutive aspect of language." The popular underlying idea seems to be that language is an instrument for communication and thus the meanings of words must be publicly accessible (Child 1994: 33). Even when we accept this idea, it is not enough to make the case that beliefs, even in the case of Davidson's radical interpretation, must be publicly accessible in the way that meanings are. This might be true with respect to some beliefs: for example, given that speaker-meaning is itself a kind of belief, namely, a speaker's belief about what his words mean, then such beliefs must be accessible if communication is to be possible. But it does not follow from this that all intentional states inherit the publicity of the public meanings. A Davidsonian needs to back up the claim that there are no more facts about what one believes than those that are accessible to "a fully informed interpreter". The considerations related to communication and the interdependent ascription of beliefs and meanings are not enough to prove this strong claim. Moreover, even if intentional states were public, this would not yet yield the interpretivist thesis. For interpretivism contains an additional claim about the constitution of the possession of such states by interpretability. The publicity of intentional states needs to be supplemented by such considerations.

However, there is an idea in the discussion of publicity that is important in developing an argument for interpretivism. In what follows, I try to make this explicit. Related to the foregoing discussion of behavioural criteria and the publicity of meaning is the idea that there are no factors that are hidden in principle. Let us call this the principle of manifestability.

> *The principle of manifestability*: Everything is in principle manifestable.

This is not the outmoded operationalist or behaviourist requirement that every ascription of belief should be related to some observable conditions or behaviour. Similarly, it has nothing to do with the verificationist theory of meaning, which associates all meaningful sentences constitutively with their verification conditions. I also think that it is distinct from Dummettian antirealism, which denies the existence of verification-transcendent truths (see Dummett 1978); however, on some construal, it could entail the latter. I take it that the principle of manifestability captures the more uncontroversial idea that no sense can be made of an entity whose existence does not manifest itself in any way. This is not to say that the manifestation must be manifestation to observation; some factors can be manifested only in a theory. To clarify the notion of manifestation further, it can be said that an unmanifestable entity is an entity that necessarily has no causal effects. Given the assumption that if an entity has causal powers, these powers are manifestable, we can relate the principle of manifestability to Alexander's dictum: "to be real is to have causal powers" (Kim 1993d: 202). This leads to a picture of

the world in which everything is causally connected with other things (taking 'causality' in quite a general sense), so everything is either cause or effect. The idea is that there is nothing that is isolated. As Wittgenstein memorably said, "a wheel that can be turned though nothing else moves with it, is not part of the mechanism" (Wittgenstein 1953: § 271). Applied to the present case, the thought is that there are no unmanifestable wheels.

I take it that the principle of manifestability, rather than considerations of behavioural criteria and the publicity of meaning, should be employed in devising an argument for interpretivism. It alone does not give us the interpretivist thesis, but it is one step towards it, and we should try to supplement it with some more specific considerations about intentional states and interpretability.

6.2.3 The consideration from folk psychology

This consideration is also discussed by Child (1994: 33–34). The widespread idea is that folk psychology constitutes a theory, or that it can be represented as a theory that contains constructs and posits that are defined in terms of their interrelations and *ceteris paribus* principles. Belief is one such construct whose role is fixed by its use in understanding people, including oneself, in explaining the actions of oneself and others, and so on. The argument states that if, for example, the notion of belief is indeed determined by this role, it has no aspects that are not involved in interpretation, which builds on that role. The same applies to other mental states.

I accept this consideration with certain reservations. For example, one does not have to follow common-sense functionalism and say that a belief's role in folk psychology determines the meaning of 'belief'. My subtly different claim, as outlined in Section 4.3, is rather that folk psychology provides the required interconnections that make the application of mental terms feasible.

As I have already argued in Chapter 5, interpretability itself is not part of the meaning of 'belief' or other mental terms. Such controversial claims about what is involved in the meaning of mental terms would only make the interpretivist position vulnerable to objections. What is required for interpretivism is only the claim that someone has a mental state *S* just in case this person satisfies the conditions that are required for having the state specified by the concept of *S*. To get interpretivism, of course, this claim should be supplemented with considerations about the need for interpretability. Here it becomes apparent that I propose interpretivism not as an account of what belief is, for it is whatever our concepts and folk psychology say it is. Rather, interpretivism is an account of what constitutes the possession of mental states, including beliefs.

These considerations from folk psychology are part of the case for interpretivism only if two conditions hold. First, and this is far from being uncontroversial, the nature of our mental states (and mental events, properties, processes, etc.) must indeed be fully reflected in our respective mental terms, the usage of which is regulated by folk psychology. In Chapter 4, I tried to defend such an approach and criticised the idea that our use of mental terms points towards some natural kinds, which we grasp via common-sense vocabulary only incompletely. The second condition concerns the scope of the mental and the idea is that there are no mental states for which mental specifications are lacking. For every mental way there is, there is a way to specify it. This point can be supported with the following consideration: if there are mental states such that they lack essentially a specification in mental terms, then there is no reason to classify them as *mental*.

6.2.4 The consideration from stance-dependent patterns

A final consideration that I discuss here recalls one answer to a recurrent objection to interpretivism discussed in Section 3.3. There, the charge that interpretivism involves the conflation of metaphysics with epistemology was answered by pointing out that mental patterns are such that their possession is intertwined with epistemic elements. This claim was based on the Dennettian idea that there are certain intentional patterns that can only be discerned from a special point of view, from the intentional stance. Intentional states are recognitional or interpretation-dependent entities (*entities* in the pleonastic sense) that become manifestable only to those who are in the position of the interpreter. If this is true, then an account of what it is to have such entities involves epistemic elements: mention of ascribability by a potential ascriber is ineliminable.

This need not entail that the patterns are in some sense unreal or subjective – the pattern that becomes visible to the interpreter is not an arbitrary creation or construction. The interpreter does not make it up, for she is constrained by already existing data. Still, the mental pattern becomes discernible only from the intentional stance and this discerning is not a matter of simple exposition of a pre-existing configuration. There is room for interpretive variation (see the discussion in Section 5.5). In sum, this consideration says that intentional patterns are constitutively connected with interpretation from the intentional stance.

We are now in a position to make use of the foregoing considerations. I have brought out several considerations, some of which can be worked into an argument for interpretivism. Here is one attempt.

1. Everything is manifestable. (the principle of manifestability)
2. The possession of mental states is manifestable. (from 1)

This connects the possession of mental states with the manifestability principle, and blocks peculiar hidden mental facts. The consideration from stance-dependent patterns gives us (3) and (4):

3. The possession of mental states is a stance-dependent intentional pattern.
4. The possession of stance-dependent intentional patterns constitutively involves ascribability.

It follows from (3) and (4):

5. The possession of mental states constitutively involves ascribability.

Now, let us return to the consideration from folk psychology. When we connect this consideration with (2), we get the idea that mental states manifest themselves through folk psychology. Mental terms and their usage in folk psychology fully reflect the nature of mental states and circumscribe the scope of the mental. The possession of mental states thus manifests itself through the application of mental terms. We can put forward the following entailment claim:

6. The possession of a mental state *S* by X entails that the mental term *s*, which specifies *S*, is applicable to X (via folk psychology)

To say that mental terms are applicable to one is just another way of saying that one can be described in mental terms or that mental states are ascribable to one by using mental terms. Thus we get (7):

7. The possession of a mental state *S* by X entails that the mental state *S* is ascribable to X (via folk psychology) by using the mental term *s*.

However, we need a stronger connection between possession and ascribability than entailment. What we need is a constitutive connection. Here we can rely on the consideration concerning the stance-dependent patterns, namely on (5). Given that the possession of mental states constitutively involves ascribability, (7) can be upgraded to a stronger claim:

8. The possession of a mental state *S* by X constitutively involves that the mental state *S* is ascribable to X (via folk psychology) by using the mental term *s*.

We are almost there. Instead of 'constitutive involvement', it would be clearer to talk in terms of constitutive conditions. Note that it is not claimed that we are dealing with the sole constitutive condition, for there might be further conditions (for instance, recall the requirements on canonical ascribability outlined in Chapter 5). In addition, the formulation could be made more general and the qualification concerning folk psychology could be shortened. The conclusion of the argument is thus the following:

9. A constitutive condition for X to possess mental states is their ascribability to X in folk psychological terms.

Perhaps this line of reasoning involves notions that are not precise enough to turn it into a knock-down argument for interpretivism. But it helps to explicate those lines of thought that might lead one to hold a position like interpretivism. For opponents of interpretivism, this would in turn allow a targeting of objections directly at its underlying considerations. One must be warned, though, that if this argument fails to convince its opponents, this does not amount to a rejection of interpretivism, given that it can be supported in other ways as well.

PART III

Extending the view

CHAPTER 7

Interpretivism and mental causation

The remaining part of the book is devoted to an elaboration of the version of interpretivism that I favour. Chapters 8 and 9 discuss issues in perception and self-knowledge. The present chapter extends the interpretivist account to mental causation. My aim is to argue that an interpretivist can make sense of mental causation. By 'mental causation' is usually meant the production of behaviour by mental states as well as the effects mental states can have on other mental states. Mental causation is implicated by the explanatory role that is assigned to propositional attitudes in folk psychology. The common assumption is that in order to explain someone's action, we refer to his beliefs and desires. In addition, one of the central motives for introducing the notion of content was to give it an explanatory role of some kind. When we explain someone's behaviour, we do not just cite his beliefs and desires, but they should have the right sort of contents relevant to the behaviour under consideration.

It is difficult to find a causal role for mental entities in any account of the mind. Some of these difficulties will be explained in what follows. However, it may seem that the prospects for tackling mental causation on the interpretivist account are especially bad. *Prima facie*, there seems to be a direct tension between the idea that mental states and contents have a causal role and the interpretivist understanding of the mental. On the one hand, it seems that causation requires internal states that can produce behaviour. On the other hand, it is not clear how ascribed mental states could play this causal role, since the ascription of these states seems to carry no straightforward commitments to the internal, causally efficacious items (see Child 1994: 123ff, for an exposition of this worry).

This chapter is devoted to the analysis and partial refutation of the intuition that ascribed mental states and contents have no causal role. The plan of the chapter is the following. In the first part, I try to formulate the problem in its full significance. I will begin my discussion with a well-known account of mental causation that has been proposed by a fellow interpretivist: Donald Davidson's anomalous monism. I will make explicit why we cannot rest with Davidson's solution, but there are also some features of his account that are worth keeping. Thereafter, I will review the arguments which show that some other widespread attempts to account for mental causation are not satisfactory. It seems that mental states cannot play a causal role after all. In the remaining sections of the chapter, I outline

an approach to mental causation that proceeds from the distinction between causal efficacy and causal relevance and place it in the context of the metaphysical scheme sketched in Chapter 2. In its different aspects, this approach is inspired by Child (1994) and Steward (1997), as well as by the distinctions introduced by Jackson and Pettit (1988). I hope that the resulting account helps to reconcile two *prima facie* conflicting intuitions: the conviction that the mental plays a causal role and our inability to see how this could be possible. Although this solution is not the only option available for an interpretivist – as is evident from Davidson's approach that I reject – it is one that is consistent with the general stance on the place of the mental adopted in this book.

7.1 Davidson on mental causation

In this section, I analyse Davidson's account of mental causation. I will begin the discussion with Davidson, since his views, outlined mainly in Davidson (1970, 1980a), represent the only sufficiently developed account of mental causation that is on offer in the interpretivist camp.[1] In addition, it is one of the most commonly considered views on mental causation.

In Davidson's approach, a singular causal relation takes place between two events. One event causes another event, no matter how they are described. Our statements about causal relations are extensional, that is, their truth-values do not depend on the way we choose to describe the causally related events. For example, let us take two statements:

(a) Sophia's death was caused by the accident involving her car;

(b) The death of Philip's neighbour was caused by the accident involving her red Rover.

1. Dennett has touched upon the topic of mental causation only occasionally and he has changed his mind on the issue over the years. In Dennett (1969: 35), he argued that explanation in terms of reasons is not causal. Later, his view was that real patterns are also causal patterns, for such a pattern is "a difference in the world that makes a subsequent difference testable by standard empirical methods of variable manipulation" (Dennett 1991b: 44, n. 22). In Dennett (1995: 412–422), he discusses an example of two black boxes with the purpose of showing that there are higher-order causal regularities and generalisations that cannot be explained in terms of microlevel causal processes. His general view is that mental causation seems troublesome because philosophers operate with an unduly restricted notion of causality. I agree with Dennett as far as causal explanation is concerned, but I think that the metaphysical notion of causation benefits from motivated restrictions. Dennett himself does not distinguish between causation and causal explanation (cf. Dennett 2000: 357–358).

In this example, we have two causally related events, the car accident and Sophia's death. Let us assume that the description 'Philip's neighbour' applies to Sophia and that her car is a red Rover. Davidson's view is that since the causal relation between these events holds regardless of how they are referred to, (a) and (b) must be true if either of them is true.

In Davidson's account, causation has a nomological character. This is to say that for every causal relation, there is a law[2] that backs up this relation. This is the point where we can spell out a difference between singular causal relations and laws. Unlike causal relations, laws depend on descriptions – namely, descriptions that specify the types of events to which a given law relates (see Davidson 1993: 6). Davidson has always been keen to point out that since there are many ways of describing a given event, we may not know what law makes a given causal relation hold. What the nomological account says, then, is that in every case of causation there *is* such a law even if its form and content remain unknown: "if a singular causal claim is true, there is a law that backs it, and we can know this without knowing what the law is" (Davidson 1995: 264). In other words, every causal relation picked out by a true singular causal statement "*c* causes *e*" instantiates a law. This is not to say that every description of a cause and effect by which we pick out causally related effects is suitable for the formulation of the law.

Davidson assumes that the laws backing up causal relations are specified in physical terms and that they are strict and deterministic. He makes the controversial point that there are no such laws containing mental descriptions. How can mental events have a causal role, and causation depend on laws, if there are no laws about mental events? Davidson answers this problem with his anomalous monism. Thus, Davidson famously begins his "Mental events" by outlining three principles that seem mutually inconsistent and argues that they are consistent on his view of the mental and the physical. These principles are the following (Davidson 1970: 80–81):

1. There are psychophysical causal relations, that is, causal relations between mental events and physical events: "The Principle of Causal Interaction".
2. For every causal relation, there is a strict deterministic law that backs up this relation: "The Principle of the Nomological Character of Causality".
3. No strict deterministic laws exist for predicting and explaining mental events: "The Anomalism of the Mental".

2. There can be various constraints on what sort of laws are allowed to back up causal relations. Davidson (1970: 81) maintains that the laws under discussion must be strict and deterministic. Fodor (1990: 152), on the contrary, opts for non-strict laws whose *ceteris paribus* conditions are satisfied.

We can assume (1) for the purposes of the discussion, and the nomological character of causality as presumed in (2) has already been introduced. (3) is the claim that the mental has an anomalous character: mental predicates are not involved in strict psychophysical laws. In all such laws, the terms that pick out events are purely physical. According to Davidson, however, this does not exclude events *c* and/or *e* which instantiate the law *also* being mental, that is, also satisfying mental predicates. The very same event that is involved in the causal relation can be picked out both by physical and mental predicates and then it is both physical and mental. In other terms, since the mental event is token-identical with the event that is causally related to another event, it can have a causal role.

It seems that in this model, there is no good way to make it clear *in virtue of what* one event causes another. For Davidson, an event is a particular happening that is not further analysed into elements. For example, my throwing a stone. The existence of this event does not depend on the way one chooses to describe the happening. If my throwing the stone breaks a window, we can say that one event caused another event – the breaking of the window. What Davidson's account is unable to say, however, is *what is there* in the first event *to cause* the second event. If one views events as unstructured particulars, there does not seem to be any good way to answer this question.

However, one might try the following line of response. Causal relations are backed up by laws. In the law, the causal relation is depicted under a certain description by using certain predicates. One option would thus be to say that the first event caused another in virtue of its properties that are picked out by predicates that enter into the formulation of a law. Then we seem to be faced with a choice. Either the causal relation between particular events remains a mystery, since we are unable to say in virtue of what one event causes the other, or we claim that this happens due to some set of causally efficacious properties, namely, in virtue of those properties that are picked out by predicates involved in the causal law. The first option is unsatisfactory, and would lead one to abandon Davidson's account of events. The second option has a disturbing consequence that can also lead to the same result. The consequence is that only those properties are causally efficacious that enter into causal laws and other properties of an event do not cause anything. For example, if Davidson was right, we could not truly say that since the kitchen table has a sharp corner, I scratched my thigh by bumping into it, for there are no laws of nature that mention the corners of tables and hurt thighs. Laws of nature mention more basic events. If we keep in mind Davidson's demand for strict laws, the causal status of nonfundamental properties becomes even more shaky.[3]

---

3. Another problem with the nomological account is that it opens up the possibility for scepticism about our knowledge of causality. Since we may not know which laws back up the causal relation, as Davidson implies, and if the causal properties of an event are those implicated in

The latter line of thought is a version of the standard charge that Davidson's approach leads to epiphenomenalism about the mental. This objection is worth recalling. As already pointed out, Davidson holds that causal relations are backed up by strict deterministic laws. He concedes that mental events stand in causal relations, but at the same time, he denies that there are any strict psychophysical laws. If we presume that only those events' properties whose descriptions enter into the formulation of a law are causally efficacious, and there are no strict laws that include mental predicates, then this would leave mental properties of an event epiphenomenal (Honderich 1982; Sosa 1984; Kim 1994: 246).

Davidson's (1993) answer to this objection was to point out that, on his view, events are not causally related in virtue of their properties. Causal relations can take place only between particular events. If that is conceded, then it seems that one cannot distinguish between causally efficacious and non-efficacious properties, since an event is causally relevant as a whole, irrespective of whether it has a mental or a physical description.

This reply does not provide a satisfying answer to the objection, however. Davidson denies the claim that an event causes another event in virtue of some of its properties, since he assumes that this commits one to the view that an event's having a certain property caused its effect. In that case, an event having a property, which is a state of affairs, would stand in the causal relation, whereas Davidson wants to maintain that the causal relata are events. But the locution that there is something about an event that leads it to play the causal role, namely its properties, does not imply this. One can hold that causal relata are events and still maintain that an event caused some effect in virtue of having a certain property (see McLaughlin 1993: 33–34). But if the question 'In virtue of which properties does one event cause another event?' is allowed, then Davidson's account is still open to the charge of epiphenomenalism. For it does follow on Davidson's view that an event's causal powers are determined by the laws it instantiates, and the laws depend solely on events' physical properties, because there are no strict psychophysical laws. In this sense, events are causally related in virtue of some of their properties, namely those that make them instantiate strict laws. Davidson succeeded in avoiding token-epiphenomenalism by assuming that the very same token event can have mental and physical descriptions and this token event itself does not turn out to be epiphenomenal. But this was not the target of the original criticisms. His opponents were worried that by avoiding token-epiphenomenalism, Davidson succumbs inadvertently to type-epiphenomenalism. Mental events as a type would be deprived of causal power, since they are not involved in (strict) laws.

---

the law, we do not know what it is about an event that leads it to have such and such effects. The old Humean error theory of causal talk has popped up again, under a new guise.

Even if one accepts that the previous point that causal relata are events does not disqualify the 'in virtue of' question, one may attempt to defend the Davidsonian account by rejecting the view that one event causes another in virtue of having properties. After all, recall that Davidson talks about predicates, not properties. The view that takes events to have properties or views events as property exemplifications at times á la Kim (1976) is a different account of events and it is not the one Davidson accepts. I think that such a rejoinder would bring us back to the beginning – we still have no sensible account of the causal relation and we still have no explication of what it is about an event that leads it to play its causal role (McLaughlin 1993: 32; Kim 1993a: 22). If one takes mental properties to be pleonastic, or acknowledges only mental predicates, then one still owes an account of their causal role. The mere declaration that token events are causal and that causal relations hold between events "no matter how they are described" (Davidson 1993: 6) fails to secure any causal role for mental types.

Finally, Davidson's reliance on the nomological approach to causation invites some doubts. What is the status of the laws that back up causal relations? One has to distinguish between the law statements and the laws as relations in nature. It is reasonable to assume that the nomological account of causation presumes objectivism about the laws, that is, the view that there are indeed such objective relations in nature.[4] An alternative would be agnosticism about objective law-relations that accepts only law statements. This view remains sceptical about the objective success of law statements, and allows the possibility that our formulations may be laden with our interests and constrained by the limits of our abilities. It is compatible with the claim that due to our conceptual structure, we are bound to formulate causal laws and employ causal categories, though there is no such causation in the world. However, a Davidsonian who wants to base singular causal relations on nomological connections cannot take this line, for in that case one would attempt to reduce natural causal relations to cognitive relations (in the form of law statements). Therefore, I take it that the proponent of the nomological approach would also be committed to objectivism about the laws. And this opens up another task for the proponent of this approach: in order to accept the nomological account, one has to give a considerable defence of the objectivist picture of laws. For it is far from obvious that the laws of nature exist as objective relations, waiting to be discovered. Perhaps the nature of nature itself is singularist and does not consist in the holding of general relations. This leads us to the third option,

4. It is not altogether obvious whether Davidson takes this line, but see his remarks on prelinguistic infants acting as if they have learned laws and his claim that conceptualised laws build on such unconceptualised foundations (Davidson 1995: 275).

which is to reverse the order of reduction. In this view, singular causal relations are more primitive than causal laws, or at least not reducible to the latter. The choice between these options leads us deep into metaphysics and too far from the more immediate concerns of the present project. Concerning the latter, it is important to point out that the nomological approach rests on presumptions that are far from obvious. The stance on causation that I prefer and outline in the final parts of this chapter does not depend on such controversial views on the place of laws in nature.

In sum, we can say that the only sufficiently developed account of mental causation presented by an interpretivist is not free of problems. First, Davidson's account either ends up in epiphenomenalism or it leaves the nature of the causal relation mysterious. In order to explain in virtue of what one event causes another, one has to single out a set of causally efficacious properties of an event. If only those properties play a causal role that are mentioned in strict laws, and since Davidson holds that there are no strict psychophysical laws, then mental properties turn out to be epiphenomenal. This consequence cannot be resisted with the claim that events are efficacious as a whole, since that leaves the causal relation mysterious. The second problem with the nomological account is that it backs up causal relations with objective law-relations, but it is not clear how exactly the laws support causal relations. In addition, this presumes objectivism about the laws, a view that requires a separate defence.

This suggests that an alternative approach to mental causation is needed, which will be explored in Sections 7.3–7.5. Meanwhile, I shall continue to emphasize the need for an alternative account by discussing some contemporary approaches to mental causation and attempting to show that they fare no better than the account just discussed.

7.2 Troubles with the causal role of the mental

The topic of mental causation has proven to be extremely difficult to handle. In this section, I shall intensify the sense of the predicament by showing that various attempts to solve the problem of mental causation actually fail. The problem of mental causation is the task of finding some sort of causal role for the mental in the overall scheme of things without compromising our ordinary conception of mental causation. According to the ordinary understanding, we want to say that I opened the fridge because I wanted to drink and believed that I would find something to alleviate my thirst in the fridge. In other words, according to our ordinary conception, we want to say that my beliefs and desires caused the opening of the fridge by me.

As already pointed out, there are no laws that relate opening the fridge and having certain beliefs and desires. If we assume that there are indeed no such laws, then, as argued in the previous section, this leads to epiphenomenalism concerning the mental. Opening the fridge would turn out to be caused by some other properties of an event: those that would be mentioned in a suitable law. What I suggest for now is dropping the assumption that there is an individuating relation between causation and laws. Let us investigate other possible ways of accounting for the causal relevance of the mental.

The central notion in one alternative account is the supervenience of mental properties on physical properties.[5] I discussed the notion and its different readings in Chapter 2. What I want to concentrate here is the use of supervenience for securing some causal role for mental properties. Although there are several people who have tried to explain mental causation in terms of supervenience (e.g., Kim 1984b; Fodor 1987), we begin by taking Davidson as an example again.[6] In one of his papers, he says that "if supervenience holds, psychological properties make a difference to the causal relations of an event, for they matter to the physical properties, and physical properties matter to causal relations" (Davidson 1993: 14). To say that mental properties supervene on physical properties is to say that if events differ in their mental properties, they must differ in their physical properties as well. Davidson's idea is presumably that if those physical properties on which mental properties supervene are causally efficacious, then the difference in the supervenient mental properties of an event means that there must have been a related difference in physical properties. In this indirect manner, mental properties are also causally relevant.

But can supervenience be used to ground mental causation? Although the details of the story obviously vary, the basic idea is this: physical properties are causally efficacious (given some non-nomological notion of causal efficacy). Since there cannot be a change in mental properties without a corresponding change in physical properties, mental properties thereby inherit the causal efficacy of the latter.

This is the idea I want to resist here. Supervenience is a symptom of some more basic relation on which it is grounded. We can detect a relation like supervenience between *M* properties and *P* properties in case *M* and *P* are one and the same property, or if instances of *M* properties are constituted by or composed

5. Although it was also used by Davidson (1993) as one means of avoiding the epiphenomenalism of mental events, I call it the alternative manner since it does not presume the nomological account.

6. Kim has changed his mind about supervenient causation in more recent work (see Kim 1993b).

of *P* instances, or if *P* instances are causally sufficient for *M* instances, etc. (Kim 1993c: 167; Heil 1998b: 150–151.) In all these cases, there is a deeper explanation for why supervenience holds. If one prefers to take supervenience as a primitive relation, then one has to distinguish it from the covariance of properties. For otherwise this account is prone to the following objection: if mental properties inherit causal powers by supervening with causally efficacious properties, then this would also hold for all properties that covary with causally efficacious properties. But surely one would not be inclined to claim that all such covarying properties are causally efficacious. For example, one might notice in a restaurant that the disappearance of food from plates covaries with the movements of mouths. One might also notice that speaking covaries with the movements of mouths. There can be no speaking without the movement of a mouth. Assume now that movements of mouths cause the disappearance of food. By similar reasoning, we would be inclined to say that speaking, which covaries with movements of mouths, causes the disappearance of food. But obviously this is not the case. Those who want to account for mental causation as supervenient causation can resist this objection if they provide an account of what grounds the supervenience relation. But then it might turn out that there is no room for inherently mental causation in this picture.

There is a widely discussed and, to my mind, convincing argument against the assumption that causal powers can be transmitted by the supervenience relation.[7] This is the causal exclusion argument by Jaegwon Kim. Its more specific wording and details may vary, but, in a nutshell, the idea is the following: if mental properties supervene on physical properties, then it is their supervenience base properties that do all the causal work, given the sensible principle that no event (in the sense of a property exemplification at a certain time) has more than one cause (proviso certain genuine cases of overdetermination). Let us review this argument more closely. I will rely on its recent rendition in Kim (2003a, 2005). Kim presents the argument in two versions; only the simpler variant is given here. The argument begins with the assumptions of mental-to-mental causation and the strong supervenience of the mental on the physical.[8] If we denote mental properties as '*M*' and physical properties as '*P*', we can state the first two claims:

7. For the purposes of this discussion, we should assume mental and physical properties in the inflationary sense, as is common among those who propound mental supervenient causation. In the next section, I will discuss the issue of mental causation from the perspective in which there are mental properties only in the pleonastic sense.

8. The supervenience at issue must be strong supervenience, for weak and global supervenience are not suitable for the transmission of causal powers (Kim 1993b: 361).

1. An instance of M_1 causes an instance of M_2.
2. M_2 supervenes on P_2.

Kim (2003a: 155) notes that these two assumptions create a tension. What makes M_2 instantiated? Its supervenience base is present, and that would have necessitated the instantiation of M_2 even if the instance of M_1 had not been present. A way to resolve this tension is to make the supervenience base of M_2 causally dependent on M_1.

3. An instance of M_1 caused an instance of M_2 through causing the instantiation of M_2-s supervenience base P_2.

Kim's conclusion to the first part of the argument is that mental-to-mental causation entails downward causation, that is, the causing of P_2 by M_1. This gives us (4):

4. The instance of M_1 causes the instance of P_2.

Given the principle of the closedness of the physical, which states that if a physical event has a cause at a given time, it has a physical cause, we may say:

5. The instance of P_2 has at the same time a physical cause, the instance of P_1.

Next, Kim brings in another crucial premise in the argument – the irreducibility premise. This premise states that mental properties are not reducible to physical properties and are not identical to them.

6. The instance of M_1 is not identical to the instance of P_1.

It follows that P_2 is caused by two distinct events (Kim 2003a: 158):

7. Hence, the instantiation of P_2 has two separate and distinct causes, the instance of M_1 and the instance of P_1, and there is no causal overdetermination.

Kim rejects the option that the mental and the physical property are distinct overdetermining causes of P_2. Sometimes unproblematic cases of overdetermination occur, like two stones hitting the window simultaneously, but in the case of the relation between the mental and the physical, we would not have similar overdetermination, since the mental causal chain is not independent and separate from the physical causal chain, for, as presumed, the mental supervenes on the physical.

Given the exclusion principle that "no single event can have more than one sufficient cause occurring at any given time" (Kim 2003a: 157), one is forced to choose between two separate causes and the closedness of the physical excludes mental causation.

8. Either the instance of M_1 or the instance of P_1 must be excluded. The instance of M_1 is excluded and the instance of P_1 is the cause of the instantiation of P_2.

The general lesson from this argument is that the principles of exclusion and closedness of the physical leave no room for the causal role of supervening properties. Both principles are more plausible than the attempt to secure a causal role for mental properties by assuming supervenience. Consequently, one should seek alternative ways of understanding mental causation.[9]

Another, but only slightly different attempt to secure the causal status of the mental is the role functionalist account in which higher-order mental properties inherit their causal powers from lower-order physical properties that realise these mental properties. However, this move is as incapable of escaping from epiphenomenalism as the previously discussed accounts. All causal powers in play in such cases belong to physical realisers. Similar Kimian considerations point to the conclusion that if higher-order properties are causal only due to some lower-order properties, then higher-order properties are not causal at all. (Even worse, as argued in Chapter 2, it is hard to see how they can be separate properties at all.) Assume for example that I was nourished because I ate. Assume that due to this relation, the act of eating has a higher-order property of being nourishing. It will not do to say that I was nourished because of the nourishing property of my eating. That would be a prime example of *virtus dormitiva*. What the role functionalist is covertly claiming is that I was nourished *because of* some lower-order property of mine that is related to the higher-order property of being nourishing. And this is just to say that this lower-order property is the genuine cause of my becoming nourished.

Kim's exclusion arguments have generated a small industry of discussion, and not all of it can be treated in this chapter. I will mention only the most pressing issue. There is an objection to the exclusion argument that is known as the generalisation objection. It is said that the exclusion argument applies to all properties that supervene on more basic properties. We may assume that the world consists of layers of such supervening properties, so that the properties at a higher layer supervene on properties of the lower layer. If the exclusion argument is on the right track, then supervening properties do not have separate causal powers. If we take a human being under the view, then it can be divided into the following

9. Kim's own solution is to postulate species-specific type identities between mental and physical properties (see Kim 1998 and my discussion in Chapter 2). In his view, this is the only way to make mental properties play any causal role at all. However, due to the identification of mental properties with their realisers, it is questionable whether this is mental causation anymore. Kim (1993b: 366) himself acknowledges that when we take this road, we "may lose what makes the mental distinctively mental".

layers: at a relatively high layer, there are properties that are specified in terms of their functions. The next layer consists of neural properties of the central nervous system. Then there are the electro-chemical properties of neurones. Then there is a layer of microphysical properties of atoms, electrons, etc. We can dig even deeper and move on until we reach quantum properties. And so forth. It is not even clear that there is a lower limit somewhere that can be regarded as the absolutely fundamental layer. If the properties at higher layers supervene on the properties of lower layers, etc., or if they are higher-order properties with respect to the lower-order properties, and if all higher-order supervening properties are epiphenomena, then epiphenomenalism generalises through all layers. The causal powers of properties at each layer derive from the causal powers of lower-layer properties. Only the properties at the fundamental layer would then have causal powers. It would thus follow that the real causal powers are located somewhere in microphysics and that talk about biological, chemical or neural causes is just a misleading manner of speaking. And if there is no fundamental layer, then causal powers drain out of the world altogether (see Block 2003). As Kim (1998: 81) puts it: "It looks as though if there is no bottom level in this picture, causal powers would drain away into a bottomless pit and there wouldn't be any causation *anywhere*!" Although there are philosophers who seem to embrace such a consequence[10], it would be better to resist it, while keeping the exclusion argument. This can be done by devising an ontology in which the relation between mental and physical entities differs from the relations between natural entities at different layers. In fact, there are two tasks here. One is to give some ontological ground for the aforementioned disanalogy. This would allow us to use the exclusion argument against mental properties, if they are conceived as supervening on higher-order properties. Another task is to stop the drainage of causal powers; for this, one has to show how natural entities can be composed of more fundamental natural entities and still have separate causal powers. Since my reply to the generalisation worry requires some stage-setting, I will postpone the discussion of this rejoinder until the final section.

It was argued that supervenience and inheriting causal powers from lower layers do not help to secure a causal role for the mind. Thus far, the putative fact

---

10. Kim (1998: 77–78) presents some examples. Burge (1993: 102) takes this as a sign not to worry about mental causation in particular. Van Gulick (1992: 325) says directly: "If the only sense in which intentional properties are epiphenomenal is a sense in which chemical and geological properties are also epiphenomenal, need we have any real concern about their status: they seem to be in the best company and no one seems worried about the causal status of chemical properties." Kim (1998: 78) retorts that "this is like being told that we shouldn't worry about, say, being depressed because everyone else has the same problem."

of mental causation appears mysterious. For how could mental events cause anything, if their causal powers derive from either physical events or lower-order properties? Moreover, the very notion of deriving causal power is suspect, since it is just another way of saying that the causal efficacy belongs somewhere else. I do not make the strong claim that no other account of mental causation is feasible, but the fact is that as yet we lack a single uncontroversial account of mental causation. In the remaining part of this chapter, I present an ascriptivist view on the issue and show how it is nevertheless possible to secure some causal role to ascribed contents and mental states which is not prone to the aforementioned failures.

7.3 Causal explanation and causation

I will begin to tackle the problem of mental causation by implementing two distinctions. One will be drawn between causation and causal explanation, and this will be supplemented with another distinction between causal efficacy and causal relevance. This indicates a version of the "divide and conquer" strategy: my claim will be that there is mental causation on one side of the division but not on the other. But let us begin with some clarification of the notion of causation. When philosophers talk about causation, they always tend to generalise. There are various processes taking place in the world, different events happening, and in the philosophical jargon, we group them together under the heading of 'cause' and 'effect'. Where one thing pushes another, we generalise, and talk about one thing being causally efficacious in making the other thing move. This can be seen as supporting the view that there is no such distinct and separate process as causation. In essence, this is the point made by Strawson (1985: 120), who wrote that "there is no single natural relation which is detectable as such in the particular case, which holds between distinct events or conditions and which is identifiable as the causal relation". When we use the term 'causation', Strawson suggests that we use it as a general name applied to different processes, each of which can also be called by some more specific name.[11] This should remind us that to call those processes that bring something about 'causal' is to engage in a more or less convenient generalisation. Especially in philosophical discussion, it

11. In fact, Strawson (1985: 135) expresses it with a more Kantian flavour when he notes that 'cause' is "the name of a general categorical notion". Whether causality is a category that precedes experience (see also Strawson 1985: 128) or not does not matter in the present context, in which the important point is that 'cause' is a general notion that covers various kinds of relations. For a brief discussion of the generality of the term 'cause' in comparison with ordinary causal verbs, see Anscombe (1981: 137).

is more convenient to talk about cause and effect than to mention specific ways of bringing about events. Sometimes, however, this risks obscuring the fact that as a general notion, it can be applied to two different affairs, namely to causal relations and to causal explanations.

The distinction between causation and causal explanation was already drawn by Davidson (1967a: 697). As already mentioned, in his account, causal relations hold between events independently from our ways of describing them, whereas the validity of the laws used in causal explanations depends on the vocabulary that is being used. Although I do not think that every explanation states laws, the general distinction is worth keeping and this will be presumed in the following discussion.

Some might doubt whether the distinction between causation and causal explanation really makes sense. To say that there are causal relations in nature is to make a strong metaphysical statement and some may find such a priori metaphysical claims too optimistic about our epistemic status or too naïve given that causation seems to have no established place in contemporary physics (see Norton 2003). Bertrand Russell's remark about causation is characteristic of the latter attitude: "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." (Russell 1913: 1.) While physics is often taken as the yardstick of what really exists, some prefer to talk about causality only in the context of causal explanation and see no need for a separate metaphysical account of causation. Thus Baker (2001: 185) states: "There are no metaphysical constraints on causes. Causes are what successful causal explanations cite."

If no metaphysical account of causation is to be had, then the problem of mental causation transforms into a question about the explanatory role of mental phenomena. There would thus still remain similar overdetermination problems in the case of conflicting explanations. Kim (1995), for example, has presented the causal exclusion problem as being equally problematic for both causation and causal explanation. In the latter case, and this is a point to which we return soon, the worry is that the physiological explanation excludes the intentional explanation. In this regard, one does not make much progress with respect to the mental causation problem if one rejects the metaphysical notion of causation. On the contrary, the discussion in this chapter hopefully shows that drawing the line between causation and causal explanation brings us a step closer to the solution of the problem.

In addition, the distinction between causation and causal explanation can be made more palatable to philosophers who share Russell's attitude. Namely, it is consistent with this distinction to hold that the relation of causation becomes decomposed on further analysis. For example, talk about things having causal

powers leaves open the possibility that, in the final analysis, there might be an account of causal powers in some other terms that make no reference to powers or causation. This does not show that there are no causal powers. The ultimate analysis must explain what there is in an object, in its properties and circumstances, that allows the object to have an effect on other objects or conditions. We may presume that talk about causal powers and causal relations is an approximation of talk about such various singular processes in nature. If these are real, causation is real. The alternative would be to espouse a kind of antirealism about causation and to deny that causation is a real relation in nature. But given the way of understanding causation just proposed – as a placeholder for various relations that really take place – the denial of the reality of causation cannot be acceptable, since it has the self-defeating consequence that no processes or happenings are ever brought about in the world. I hope that such an interpretation of the notion of causation makes it fairly uncontroversial and that it might also be accepted by those who otherwise prefer to talk only about explanation.

It is one thing to distinguish between causation and causal explanation and it is another thing to explicate the basis of this division. Strawson (1985: 115) provides a good starting point in this respect (for further discussions, see Child 1994: 100–110 and Steward 1997: 165–186, who also proceed from the Strawsonian starting-point). According to Strawson, causation is a natural relation[12], whereas causal explanation is a rational relation. The first one exists in nature, taking place mainly between events; the other is an explanatory relation between "facts or truths". Both of these relations are connected with causation insofar the very same term 'cause' is used. They are both related to the same categorical notion, which, in Strawson's words, is usually referred to "in connection with the explanation of particular circumstances and the discovery of general mechanisms of production of general types of effect" (Strawson 1985: 135). On this view of the causation vs. causal explanation distinction, the difference between these relations originates from their locations and from the nature of the relata. The causal relation is located in the natural world and it relates particulars, while the explanatory relation is located in our minds and it relates facts (Strawson 1985: 117). Of course, in providing explanations, we can refer to natural particulars, but particulars themselves do not partake in explanatory relations. On this view, that role is played by their abstract kin, namely, facts. However, for present purposes, we do not need to lay strict constraints on ontology. When one thinks about the lively discussions in metaphysics over the right sort of relata for a causal relation, it

12. Or, we might say in view of Strawson's remarks (cited above) that by 'causation' we refer to all those various natural relations that consist in bringing about effects. I will omit this qualification in what follows.

seems premature to declare without extensive argumentation that only events can be causally related and not states of affairs (Armstrong 1997), tropes (Campbell 1990) or facta (Mellor 1995: 162). However, this does not fault the central idea – that causal relations occur in nature and causal explanation is a matter of rational relations. Without committing ourselves to any particular ontology of causation, we can say that causation relates entities that belong to the natural basis. It is a further task to work out an ontology of those natural entities that would enable us to understand why they bring about their effects.

Making the distinction between causation and causal explanation requires that we explicate the relation between them. The question about the their interrelation can be separated into two sub-questions: 'What makes an explanation causal?' and 'What makes a causal explanation true or acceptable?'. Here, I will devote more space to the first question. In respect of the issue of what makes an explanation true, our present topic does not demand an idiosyncratic account. If the question of the truth of our explanations sounds ambitious, we can ask what makes the explanation acceptable and here we can rely on whatever constraints philosophers of science have set for the acceptability of an explanation. Given the distinction between causation and causal explanation, the first question is more pressing.

A forthright answer to this question would be that a *causal* explanation is one where the explanatory relation mirrors the causal relation. Some might tell the same story about the truth of the explanation. For example, Child considers one such account that does not distinguish these two questions:

> The general idea, then, is that the truth (or acceptability) of a causal explanation rests on the presence of appropriate relations of causation. And a natural thought would be to put the point in the following way: a causal explanation is one whose explanatory power depends on the assumption that there are events mentioned, or pointed to, in the *explanans* and *explanandum* sentences, between which the natural relation of causation obtains; and whose truth (or acceptability) requires that that relation does indeed obtain. (Child 1994: 102)

Here, the order of validation goes from explanation to causation: if we have a piece of explanation, then it is considered acceptable if there are corresponding *causal relations*. However, if this is presented as the *sole* account of explanation, the answer will not do. The problem with it is that it privileges events as causal relata, whereas an explanation can also point to conditions or property instantiations, etc. Even if we make the appropriate amendments on this score, there still remain problems. One worry is that the causal relation is extensional and would make true any applicable description of the causal relata. Explanation, by contrast, depends on the terms involved. For example, the claim 'The event that I described in Tuesday's entry of my diary was caused by the event mentioned in my diary two

days earlier' is a causal explanation, since it involves the notion of causing, but it is completely uninformative, for it does not pick out events in the required way.[13] But the claim itself is true, presuming that the events mentioned in the respective pages of my diary were indeed causally related. When applied to mental causation, the problems with this account become even more evident. Imagine that we have two causal explanations, one of them containing mental predicates, while the other does not. Both of them pick out the very same set of causally related entities. We cannot choose between these explanations using the above criterion since, at the level of causal relations, the same entities and their interrelation correspond to these explanations. According to this criterion, however, both explanations would have equal explanatory power, for the same causal relation is in place in both cases. Then we have two conflicting explanations for the occurrence of the same effect and no other criteria for preferring one to the other. This also shows that there must be additional conditions for the adequacy of an explanation.

It may thus look as if, insofar as we want to view several functional and higher-order explanations as causal, we should drop the requirement that an adequate causal explanation depends on there being a corresponding causal relation. Then it would not be the case that the causal explanation is *causal* in virtue of there being a matching causal relation. There is indeed another view on the relationship between causation and causal explanation and this has already been introduced. This is the Strawsonian suggestion that our usage of causal notions indicates the application of a general category of causation (see also Child (1994: 109) and Steward (1997: 175–186)). On that view, the most that one would find common for various causal explanations is that they involve causal vocabulary. This gives us a preliminary understanding of the notion of causation that can be applied to demarcate a certain range of explanations and it is thus not the case that we can specify the notion of causal explanation only through the antecedently grasped idea of a causal relation. However, even in this approach we have to give some account of why we tend to apply causal vocabulary in certain situations. Otherwise, the explanation in causal terms would acquire a completely arbitrary character.

We should try to steer the middle course between these accounts.[14] One can admit that we cannot expect that in every case of causal explanation, the *explanans*

13. Both Davidson (1963: 698) and Child (1994: 101) give a similar example.

14. Child (1994) and Steward (1997) both seem to prefer the causation as a general category view, although Child (1994: 109–110) also wants to keep the assumption that causal explanation presumes the corresponding causal relation between events. Steward rejects the more specific view that causal explanation presumes the presence of a natural causal relation between *particulars*. In her account, explanations can also be underpinned by counterfactually supported natural relations between facts (see Steward 1997: 184–185).

and the *explanandum* are in one-to-one correspondence with causally related entities. One can also admit that the occurrence of a causal relation is not the primary reason why we take some explanations to be causal explanations. But all this does not show that in a true case of causal explanation there is no causal relation in place. Successful causal explanations entail the existence of some causal relation even if the latter does not set constraints on the adequacy of explanations and the explanatory relation does not mirror the causal relation. This approach is compatible with the claim that the nature of the elements of the causal relation cannot be read off from the explanation, especially when we are dealing with a functional explanation or with a complex explanation that involves several variables. From the perspective of such a higher-order explanation, the metaphysical claim about the existence of some causal relations does not even carry much explanatory weight. But from the metaphysical point of view, acknowledging such relations is important in securing realism about causation.

7.4 Causal efficacy and causal relevance

Thus far, we have adopted the distinction between causal relations and causal explanation and admitted that a causal explanation need not mirror (that is, its *explanans* and *explanandum* need not stand in one-to-one correspondence with) the respective causal relation. However, it could still be true that when we are in a position to provide a causal explanation, there is indeed some causal relation in place. It is just that the existence of such relations does little to constrain the explanatory adequacy of our causal explanations. In what follows, I attempt to elaborate on this distinction by plugging it into the distinction between causal efficacy and causal relevance, as this was first proposed by Frank Jackson and Philip Pettit. In doing so, I have been inspired by Helen Steward (1997), who elaborates her account through this distinction, although our approaches diverge in the end.[15] My eventual aim is to make use of these distinctions in giving an account of mental causation that is compatible with interpretivism. I will begin this section with giving an outline of the Jackson-Pettit distinction between efficacy and relevance and then I will adapt it to my presuppositions.

15. There is no space for a proper discussion of Steward's approach here, but I should mention some differences between us. For instance, I map the relevance-efficacy distinction exclusively onto the distinction between pleonastic and natural entities, while Steward maps it onto the distinction between facts and particulars and she allows facts to be causally relevant also to particulars. I also do not share Steward's naturalism about facts and about causal relevance, since this tends to obscure important metaphysical distinctions between efficacy and relevance as well as between pleonastic and natural entities. My contention is that the causal relevance relations between facts ultimately presume some efficacious ground.

In a series of papers, Jackson and Pettit (1988, 1990a, 1990b) distinguish two kinds of explanation. With the aim of explaining how dispositional and functional properties can play a causal role, they make a distinction between the process explanation and the program explanation. The process explanation points out which property-instances are *causally efficacious* in the process of producing the effect. Jackson and Pettit proceed from the standard assumption that in the case of dispositional properties, the categorical property instances are causally efficacious and in the case of functional properties, the efficacy belongs to the realising property instantiations. They concede that due to the exclusion issue, dispositional and functional property instances cannot be causally efficacious. However, they can still be *causally relevant*. The explanations given in terms of these higher-order (dispositional and functional) properties point out the properties that are causally relevant to their effects. They call such explanations "program explanations", since they involve a case of one property "programming for" other properties. In other words, the instantiation of a dispositional or a functional property entails the instantiation of the causally efficacious property (see e.g. Jackson and Pettit 1990b: 115). Thus, when we can point to certain higher-order properties, we can use them in causal explanations, since we know that in their presence certain causally efficacious properties must be present. In this sense, referring to such properties in an explanation has causal relevance.

However, the explanatory import of causally relevant properties amounts to more than this. The causal relevance of a property depends on there being an efficacious basis, but this basis does not determine the range of explanations for which the property can be relevant. The properties that have the same efficacious basis may have different relevance and a property can have a different basis, while keeping its causal relevance. The second part of the claim is underwritten by multiple realisability. As for the first part, then, to see that different dispositions or functional properties that have the same efficacious basis can have different causal relevance, let us take a look at an example, which is provided by Jackson and Pettit (1990a: 203–205) and also discussed by Steward (1997: 190–191). Imagine a situation in which Mary dies because she stands on an aluminium ladder that is in contact with a power line. We can explain Mary's death by pointing out that the metal ladder conducts electricity. In this explanation, we cite the dispositional property of the ladder, conductivity, which is one disposition among the many that all have the same categorical and efficacious basis, namely, the behaviour of the free electrons of metal atoms. For example, the same basis also grounds the opacity and the ductility of the ladder. However, neither of them is relevant to the explanation of Mary's death. It is the electrical conductivity that we use to explain the effect. Other dispositions that are realised by the same basis are not causally relevant when they occur without the effect occuring.

Given multiple realisability, one and the same functional or dispositional property can be realised by a different basis. Jackson and Pettit make use of this feature in order to win more importance for the explanation in terms of causally relevant properties. According to them, the causal relevance of higher-order properties consists in the fact that although the realising basis may differ, the result remains the same (Jackson and Pettit 1990a: 204.) For example, the exact nature of the categorical basis plays no direct role in the explanation that Mary died because the ladder was a good conductor of electricity. Had the ladder been made of some other metal that had grounded the same disposition, Mary would still have died by electrocution. In sum, we can say with Jackson and Pettit (1990a: 205) that the disposition is causally relevant due to the fact that the actual categorical basis of this disposition was causally efficacious in bringing about the effect and the fact that if the disposition had had a different basis, it would have been causally efficacious to the same result. In this respect, program explanation is related to how things might have been. They inform us which counterfactuals are true about the case. Giving a process explanation in terms of what actually happened does not supply such information about counterfactual scenarios.

Jackson and Pettit (1988) also apply this account to the notion of mental content. They make a related distinction between intrinsic and relational properties, where the instantiations of the former are regarded as causally efficacious and the instantiations of the relational properties as causally relevant. Their claim is that "features which causally explain need not cause" (Jackson and Pettit 1988: 392). In the case of mental events, the relational and thus causally non-efficacious properties are their intentional properties, their contents. As explicated above, this does not preclude their use in program explanations.[16]

Jackson and Pettit's distinction is framed in terms of explanatory levels and property instantiations. I agree with Steward (1997: 192) that the distinction between efficacy and relevance can be given more metaphysical weight; that is, there is no need to take it merely as a distinction between types of explanation. Given the commitments made in this book (e.g., the distinction between deflationary

16. Frank Jackson does not hold this view anymore, for the reason that the idea of causally impotent but causally explanatory mental content runs contrary to our common-sense assumptions about mental causation (see Jackson 1995: 263). He gives an example: "I am watching the screen of a radar monitor, calling out changes in the position of the blips as they occur. It seems that what I call out is controlled by the content of my perceptual belief as to the movements of the blip." (Jackson 1995: 263.) I agree that this is the way common sense would explain the situation and this explanation is fine as a piece of folk psychology. This does not entail that common sense is right about the causal powers. What really happens is that the position of the blip causes some definite neural activation patterns in the visual cortex which in turn causes the characteristic activity in the area that controls verbal response.

and inflationary properties outlined in Chapter 2), this distinction cannot be applied directly to the ascription theory. However, it can be translated into its terms. Their insight that I would like to keep is the distinction between causal efficacy and relevance. I will clarify this in what follows. The fact that we can give a causal explanation indicates the presence of some causally efficacious ground. This ground can only consist of inflationary entities, for only they can be causally efficacious. This point should be fairly trivial, for only inflationary entities have enough "substance" to stand in real causal relations. Pleonastic entities are entities only in the deflationary sense and as such they are too abstract to be efficacious. The claim that only inflationary entities are causally efficacious does not yet explain the source of causal efficacy. This account should be provided separately. For example, one option would be to argue that efficacious inflationary entities must be the bearers of causal powers and then to explain the origin of such powers. But there is no need to commit ourselves here to one metaphysical theory or another. Neither do we need to commit ourselves to the claim that only properties can be efficacious. Other entities of the natural basis might turn out to be efficacious as well, depending on how the final ontology comes out. For present purposes, all that matters is the claim that causal efficacy can only belong to inflationary or, even better, natural entities.

Deflationary entities, on the other hand, can be *causally relevant*, but they cannot be causally efficacious. When we explain in terms of pleonastic entities, the explanation is often formulated in counterfactual terms and it shows our interest in explaining what would have changed, had some conditions been different. By bringing the former distinction between causation and causal explanation to bear on the present issue, we can say that the causal relation, which takes place in nature, involves causally efficacious entities. Entities that can only be causally relevant are not part of causal relations themselves; they come into the picture only in a causal explanation. However, we should not thus assume that the causation-explanation distinction matches one-to-one with the efficacy-relevance distinction. Their relationship is rather more complex: whereas the natural causal relation involves only entities that can be causally efficacious, one can also refer to efficacious entities in a causal explanation. This does not mean, of course, that every causally relevant entity we mention in a causal explanation is causally efficacious.

A causal explanation in terms of pleonastic entities indicates the presence of some natural relation. In cases where the entities at issue are trivial pleonastic counterparts of terms denoting the elements of the natural basis, the nature of this relation can be read off from the explanation. But in most other cases, where the relation between the natural basis and pleonastic properties is not so straightforward, it is not transparent which causally efficacious entities are standing in the causal relation.

7.5 The relevance of mental causation

In this section, I am finally in a position to outline an account of mental causation that builds on the distinctions made in Sections 7.3 and 7.4. I proceed from the assumption that there are no inflationary mental entities. This assumption was outlined in Chapter 2 and it has informed my interpretivist search for an alternative account of the mental in this book. All mental entities are merely pleonastic. This allows quite a straightforward stance in respect of the mental causation issue. Namely, it follows that mental entities are not causally efficacious. Since there are no mental inflationary entities, there is no causal efficacy that is inherently mental. However, mental pleonastic entities can be causally relevant. Of course, the latter claim depends on the validity of the widespread presumption that folk psychology conceives mental entities in causal terms. But I think that this assumption should be fairly uncontroversial given that much of folk psychology involves explaining action by reference to mental causes. An alternative view is that common-sense explanations are not causal explanations – instead, they are rationalisations, aiming to make sense of the action in the light of one's beliefs and desires. However, rationalisation does not exclude causal explanation. We can say both that having such and such a belief caused one to act in this characteristic way, and that citing this belief allows us to make sense of that action. In addition, as Davidson (1963: 691) famously argued, we need the assumption that reasons cause actions in order to single out the reason why one acted among other reasons that one might also have had, but which did not lead to this action. Then we can say that one's action was caused by that reason and not by the others.

Accordingly, if we grant that mental entities are involved in common-sense causal explanations, we can assign causal relevance to mental entities. We use mental vocabulary in our common-sense explanations because this corresponds better to our explanatory interests than, for example, the reference to neural events in the brain. Mental facts and properties, in the pleonastic sense of a fact and a property, are thus not epiphenomena, since something is an epiphenomenon if it is devoid of causal power for some reason, but could have it otherwise. Pleonastic facts and properties, however, do not belong to the category that can have efficacy, and therefore the worries concerning epiphenomenalism should trivially fade away.

The account in which mental entities are causally relevant but not causally efficacious also helps to reconcile a pair of seemingly conflicting intuitions. It seems that, on the one hand, mental entities must play a causal role somehow, since they are involved in causal explanations, and, on the other hand, it is difficult to see how this can be the case, given the worries about exclusion and overdetermination

as well as the intuition that only physical entities have the power to bring forth effects. But if mental entities are only relevant but not efficacious, then these intuitions are not in conflict after all.

What about the issues of causal drainage and exclusion that were discussed in Section 7.2? It should be noted that Jackson and Pettit (1990a: 209) are happy with the view that the only efficacious properties are to be found in microphysics. This means that they allow the drainage of causal powers to basic physics. If we want to stop the drainage and admit that neural entities can also have causal powers, then we should find some alternative solution. What we need is an account that shows how natural entities of varying complexity can be causally efficacious, without the consequence that their causal efficacy is derived from the entities of fundamental physics. For if we take the latter option, then the previously rehearsed exclusion argument would apply and we should accept the consequence that such natural entities as chemical processes in the brain are not causally efficacious.

One way to make sense of this would be to think of higher-layer natural macro-entities (this does not include mental entities) as composed of lower-layer natural micro-entities and the relations between them. As such, they would also have different causal powers than each of their components taken separately. The worry that higher-layer causal powers are absorbed by lower-layer powers does not apply in this case, since the causal powers of macro-entities are not identical with the powers of the micro-entities that compose them. Neither are they identical with the mereological sum of the powers of micro-entities, since this does not take into account that the relations between micro-entities can bestow a distinct causal power on the macro-entity that is composed of them. (For a related but somewhat different approach, see Kim 1998: 80–86, 116–118 and 2003a).

However, for present purposes, it is not necessary to give a final account of this issue. It is enough to know that there are ways to stop drainage if layers of natural entities are at issue. In the present context, it is more important to take a look at the exclusion issue. Recall that in Section 7.2, I argued, following Kim, that if mental properties are regarded as supervening or higher-order properties, then they are vulnerable to the exclusion argument. Distinguishing causal efficacy and relevance helps to avoid at least some charges of overdetermination. Pleonastic entities do not contest the causal status of natural entities, for they make no demands on causal efficacy.

One might also follow Kim (1995: 130) in arguing that exclusion is a problem for both causation and causal explanation. Nevertheless, it seems that, given the distinction between causal relevance and efficacy, the exclusion problem ceases to be a *metaphysical* worry. There can be no exclusion issue concerning the causal efficacy of mental entities if the latter can have no efficacy at all. It seems that it is easier to solve the explanatory conflict than the metaphysical problem, so in this

respect we have made a little progress after all. In Kim's (1995: 130) view, explanation and metaphysics are closely connected: "The only way in which I believe we understand the idea of a causal explanation involves the idea that the event invoked in a causal explanation is in reality a cause of the phenomenon to be explained." I do not think this is the only way, for as we have seen in Section 7.3, causal explanation can be connected to natural causal relations more loosely than Kim envisages. It is true that when we can give a causal explanation, there must be some respective causal relation in play, but we can understand causal explanation even if we do not know which natural events are causally related in a given case. We can be satisfied with causal explanations in counterfactual terms that involve very high-layer descriptions or we can explain in terms of networks of variables (e.g., as in Spohn 2006), in which case there are also no simple and straightforward relations between natural entities or events.

This is not to deny that there can be different causal *explanations* stated in mental and physical terms that seem to exclude each other. One way to avoid such exclusion is to argue that the mental explanation and the physical or neural explanation have different explanandums, so they are not conflicting explanations of the same effect after all. Where such a solution is not applicable, we can argue that the putative overdetermination is not vicious; that is, that the causal relevance of a pleonastic entity for the purposes of explanation depends on our interests and the issues that we want to investigate. There is no prescription that only one pleonastic cause is relevant to a given effect.[17] After all, it could also be the case that both mental and physical properties indicate the presence of one and the same efficacious basis. This does not entail that the explanation in terms of reasons is superfluous, for it gives us modal and other information that we would lack otherwise.

This leads to a question about the relationship between the causal relevance of the mental and the efficacy of the natural basis. Given the view outlined in the previous sections on the relationship between causal explanation and causal relations, giving causal explanations in mental terms need not imply token identity or other kinds of reductionism. But it is still the case that when an action occurs, some physical causal process must take place. Its explanation in folk-psychological terms does not point out the relata of this process, but it does not follow that there are no efficacious entities involved. Similarly, when we make a

---

17. Here, I am in agreement with Helen Steward, who uses the following metaphors: "facts do not compete for causal space in the same way as particulars. Facts are tolerant for other facts in the community of causal *explanantes*; they do not need to fight one another for *Lebensraum* in the way that events seem to have to do" (Steward 1997: 261). To apply these claims to the present context, replace 'facts' with 'pleonastic entities' and 'events' with 'natural entities'.

successful prediction of an action, which is confirmed by the further course of events, there must again be something causally efficacious occurring or being instantiated. An interpretivist can concede that much. On the other hand, this still tells us little. The relation between the mentalistic explanation and natural facts remains loose.[18] Canonical ascriptions that mention mental states to explain or predict acts can give rise to the hypothesis of a corresponding subpersonal level causal relation between the movement and some subpersonal surrogates of those mental states, but usually this is not enough for more detailed predictions. The existence of a causal relation that is implicated by true personal-level claims is not enough to generate a conflict with the interpretivist outlook.

The present chapter was motivated by the need to find an account of mental causation that is compatible with the ascription theory that, *prima facie*, seems to have the disturbing consequence that mental entities cannot play a causal role. I reviewed some ways of securing the causal status for the mental, including Davidson's anomalous monism, and found them inadequate. The search for an alternative account began with the distinction between causal relations and causal explanation that was then supplemented with a distinction between causal efficacy and causal relevance. Only inflationary entities can have causal efficacy, and deflationary entities can only be causally relevant in an explanation. If this is applied to mental causation, then given the denial of mental inflationary entities, there is no place for causal efficacy in the mental realm. On the other hand, pleonastic mental entities have causal relevance in various explanations. In this way, we have satisfied two seemingly conflicting intuitions: the first is the conviction that the mental plays some sort of causal role, nurtured by the fact that we explain our actions by reference to mental causes; the second is the inability to see how that could be the case. By distinguishing two kinds of causal relations that still fall under the general notion of a cause, we can make sense of both intuitions. Mental entities have a causal role in virtue of being causally relevant. The reason why it was hard to explain how mental entities can be causes is that this task was previously taken to consist in explaining how mental entities can have causal efficacy, which indeed they cannot have.

---

18. Child (1994: 112) describes a similar position as "the most relaxed view about psychophysical correlations".

CHAPTER 8

Perception

The aim of the present chapter is to explore whether perception can be treated along the lines of the ascription theory. I argue that there are no obstacles in this respect. Although there are several differences between paradigmatic propositional attitudes (e.g., beliefs and desires) and perceptual states, both of them are amenable to the ascription theory. Not attempting to provide an exhaustive overview of philosophical accounts of perception, this chapter presents a suggestion for an ascriptionist treatment of perception. I do not claim that this is the only account open to an interpretivist. Indeed, there are presumably other options even if one proceeds from the ascription theory. The account presented here is simply the one that I find most plausible given the special properties of perceptual states and contents. It should be noted that I concentrate mainly on the case of vision at the expense of other sensory modalities.

The chapter begins with a discussion of the differences between perception and belief. This is also related to the issue of the veridicality of perception, although my investigation is not specifically epistemological. After that, I will examine what is involved in the ascription of perceptual content. The chapter concludes with a section on the phenomenal features of perceptual states. My contention is that a suitably modified ascriptionist approach can also accommodate the intuition that there is something it is like to have an experience.

8.1 The specificity of perception

The ascription theory is a general theory about how mind arises in interpretation. In application to various mental types, it should consist, for every mental type that it treats, of two parts. The first part outlines the central common-sense platitudes about any given mental construct. The second part tells us how these platitudes are implicated in the interpretation. As we saw already in case of mental causation, the ascription theory also has to take a stand on a variety of philosophical issues – especially on those related to the philosophy of mind – and the option of criticising these folk platitudes should remain open. Hence, in this section, I will review some important platitudes about perception with the aim of developing a clearer view of the possible constraints on the ascription of perceptual states.

There is one aspect in which perception is unlike belief. There is an intuition that one may see something even if one is not able to say what it is that one sees. To put it another way, perception seems not to require concepts, whereas beliefs are always conceptual. Although there are some dissenters, a widespread assumption is that there are no beliefs which cannot be expressed in conceptual form. I will discuss this difference in Sections 8.4 and 8.5, where I examine the view that perception has non-conceptual content. There I also attempt to say more exactly what is involved in having a state with conceptual content. At present, let us take this as one point of difference between the perceptual and the belief-like states.

A related difference is that perceptual content seems to be encapsulated. The idea is that beliefs can influence other beliefs. By having some beliefs, one can arrive at other beliefs that are inferentially related to the former. By contrast, experiences are not generally penetrable by beliefs. Of course, in some cases one's strong expectations influence what one sees, but in normal cases one's attitudes do not influence one's perceptions. This is especially evident in case of illusions like the Müller-Lyer illusion in which horizontal lines of equal length are perceived as being of unequal length. This illusion persists even if one believes that the lines are equal – one cannot help seeing them as unequal. Such encapsulation of perception is sometimes given as a reason to treat perceptual content as non-conceptual. The idea is that perceptual content cannot be conceptual, since, if it were, it could be influenced by other conceptual contents. But it is not influenced in this way, hence it must be of some other kind.[1] This gives us another reason to think that perceptions are states of a different sort to beliefs.

The final difference to be noted here will be quite important in the following discussion. It is the difference that perceptual states have in relation to the world. In a nutshell, beliefs seem to be more easily prone to illusions than perceptions. Let me explicate this. In the case of beliefs, it is not of utmost importance that they be true with respect to the world. Of course, it is preferable to have more true beliefs than false beliefs, and having true beliefs has practical advantages. However, it is characteristically easy to entertain beliefs with contents for which there are no facts in the world. Just think of all the beliefs about fictional entities; beliefs about what would or could happen. The conceptual world of fantasy knows few limits. By contrast, one's perceiving something entails the existence of this thing or state of affairs. This sort of dependence on what there is is called *factivity*, and I discuss it more closely in the next section. The notion of factivity highlights the point that such states can be had only when there is a corresponding fact of the matter (see

1. This talk about influencing need not conflict with interpretivism and presume ascription-independent information processing, for the idea of one content influencing another can be included in the theory on which the interpretation is based.

Szabo Gendler and Hawthorne 2006: 20). Knowledge is also a factive state. One cannot know what is not the case. Similarly, one cannot see something that is not there. Illusions and hallucinations are not cases of seeing. Of course, it would be misleading to say that all perceptual states are factive, for there is a subclass of states that do not imply factivity. These are expressed by success-neutral descriptions like 'it seems', 'it looks as if', 'it appears'. I discuss such states in Section 8.3. This does not diminish the importance of factivity considerations.

I have outlined three differences between perceptions and beliefs. Not all of them are uncontroversial, but I hope to explain the troubling aspects away in the next sections. Perceptions are non-conceptual, conceptually impermeable and factive, whereas beliefs are conceptual, conceptually integrated and non-factive. These differences provide a basis for distinguishing between perceptual states and perceptual beliefs. There are philosophers who prefer to ignore this distinction and treat perception as acquiring perceptual beliefs or conceptual belief-like states (e.g., McDowell 1994; some papers in Davidson 2001c). However, their reasons for doing so are disputable. One such reason is that perceptions should justify beliefs and they cannot justify beliefs if they are states of some different kind. I discuss this argument in Section 8.6, when I defend my approach to non-conceptual content. In what follows, I assume that there are reasons for not assimilating perceptual states to perceptual beliefs and I will argue that their ascription conditions differ as well.

8.2 The factivity constraint

Given that the perceptual state differs from perceptual belief, how does this relate to the ascription theory that has been mainly worked out on the example of paradigmatic propositional attitudes? For a start, we may say, as with the latter states, the condition for one's having a perceptual state is that it can be canonically ascribed to one. It follows that no room is left for perceptions that cannot be ascribed. The full picture of the ascriptionist treatment will emerge during the following discussion. Here I focus on changes to the account that are demanded by the specificity of perception. Namely, the question is: if certain perceptual states are folk-psychologically understood as factive, does the ascription of factive states impose some special conditions on the process of ascription? In other words, are there differences in the ascription conditions for non-factive and factive states?

The factivity condition can be formulated as follows:

(F) A factive mental state with the content *p* occurs only if it is the case that *p*.

For example, John can only see that *p*, if it is the case that *p*. For a moment, let us ignore the complications arising from the difference between seeing *p* and seeing *that p*. I address these in later sections, especially in Section 8.6. I take the factivity condition as a common-sense platitude about certain kinds of mental states. Of course, common sense does not formulate this in such terms. But the platitude is still evident in common-sense claims of the following kind: 'You cannot see a rabbit, if the rabbit is not there'. Given that we idealise folk-psychology for philosophical purposes, we can say that (F) incorporates some folk platitudes.

How to include (F) to the ascription theory? My proposal is to consider the following constraint:

> (FC) The ascription of a factive mental state with the content *p* is canonical only if it is the case that *p*.

This is a condition that plays a role in deciding which ascriptions are canonical. Namely, the holding of a respective fact is a necessary condition for the canonical ascribability of a factive state. If the fact that *p* does not hold, the corresponding factive state with the content *p* is not canonically ascribable. Of course, the factivity constraint does not replace the usual conditions of canonicality: the conditions concerning revisability and the coherence of different sources are still involved. Indeed, the factivity constraint influences the outcome of the ascription through the third source, the source of stimuli, and the ascription statement has to fit with all the relevant sources. But the (FC) constraint has an influence on the kind of state that is canonically ascribable. If it is not the case that *p*, then a state of *seeing p* cannot be ascribable. Instead, a state of *seeming* that *p* or some other such non-factive state would be ascribable canonically.

The point that in the ascription of perceptual states, the object of perception plays the most important role, compared to other ascription-sources, can be made more explicit if we consider the actual ascription of factive perceptual states. When making such an ascription, it is practical to consider the stimuli affecting the subject. Let us introduce the notion of a *singularly compelling ascription*. A well thought-out ascription usually relies on a variety of sources. In the case of a singularly compelling ascription, only one source would lead to the ascription. The need to consider other sources falls away and the ascription is actually made on the basis of a single source. This fact need not make the particular ascription uncanonical, for it can be the case that other ascription-sources cohere with the proposal of the single and compelling source or that other sources have no relevance for a given ascription. If this is the case, then we say that this single source is considered by default as the primary source. However, the situation changes when discrepancy arises between the proposal made by the singularly compelling source and other possible ascription-sources. For example, it may turn out that

the person cannot see an apple, since the apple-like object in her visual field is actually made of paper or there is something wrong with her visual mechanisms as a result of which her vision is impaired. In such situations, it turns out that the actual ascriber was mistaken about the declaration of the primary source and the additional information provided by other sources would lead to the correction of this declaration. Such possibilities of error show that a source is singularly compelling for an actual ascription only in a defeasible manner. But we can still say that for certain kinds of mental states, there are singularly compelling sources for their ascription and, for factive perceptual states, this source is those stimuli that affect the subject.

8.3 The veridicality of perception

The factivity of perceptual states has to be distinguished from the issue of the veridicality of perception. Veridicality is related to the correctness of perceptual content, while factivity is a feature of a certain class of perceptual states. Of course, these issues are related: factive states are also veridical and cases of non-veridical perception have to be specified in terms of non-factive states. The difference is that whereas certain non-factive states can still be veridical, factive states cannot be non-veridical. For example, it may look as if there is a grey rabbit passing by, and indeed there is, but one cannot see that an apple is on the table in front of one while there is no apple there. We may say that perception is normally veridical, but the veridicality of perception is defeasible, for perceptual illusions and hallucinations may occur. By contrast, there is a class of perceptual states that are always factive, without any additional defeasibility conditions.

Non-factive states can be subjectively indistinguishable from factive states. They can both lead to similar beliefs. Consider the following three claims:

1. If it appears that *p* to S, then S believes that *p*
2. If S sees that *p*, then S believes that *p*
3. S sees that *p*, only if it is the case that *p*

The statement (3) is the factivity condition. (2) talks about the factive state of seeing and (1) contains the non-factive verb 'appears'. The statements (1) and (2) express common-sense platitudes about the relation between a person's perceptual appearance and his beliefs. It is part of our folk psychology that perceptions provide reasons for perceptual beliefs.[2] Both factive perceptions and

2. Cf. Spohn (1998: 31), who advocates the claim that how it looks to a person is a defeasible a priori reason for this person to assume that it is so.

non-factive perception-like states lead to perceptual beliefs. The difference is that it is necessary that in the factive case, the corresponding fact also holds. The platitude about the relation between perceptual states and beliefs plays a role in the ascription of perceptual beliefs. When we interpret someone and ascribe beliefs to him, we must also consider his perceptions. The ascription theory says that given the ascription of a perceptual state, there is a *ceteris paribus* commitment to the ascription of perceptual belief. This consideration is constrained by the principle that one has a reason to believe that *p*, if it appears to one that *p* (barring possible discrepancies).

Let us now return to the issue of veridicality. Presumably, the normal condition of perception is to be veridical. Our sense organs have evolved to detect environmental stimuli. We are built to respond accurately (that is, in a veridical way) to the stimuli that affect our organs of perception. But as we all know, there are exceptions to this rule. There are various kinds of visual illusions where we see that objects have properties that they do not have, and there are hallucinations in which we see objects that are not there. The ascription theory has resources to accommodate both types of case. One may distinguish between six possible states of affairs in this respect. They differ in respect of the veridicality of states and the correctness of ascription. Note that distinguishing some of the cases requires keeping the distinction between actual and canonical ascriptions.

Given that the fact that *p* is given to the subject A as a stimulus, the possibilities are the following:

a. *The perceptual state is veridical and it is ascribed as veridical perception*. In terms of the ascription theory: 'A sees that *p*' is canonically ascribable and it is the case that *p*. The fact that *p* obtains and A can be interpreted as seeing that *p*.
b. *The perceptual state is a veridical hallucination and it is ascribed as veridical perception*. It is actually ascribed that A sees that *p*, and it is the case that *p*. However, although A hallucinates that *p*, accidentally the *p* situation obtains. It is not canonically ascribable that A sees that *p*, but such an ascription is made by mistake. What is canonically ascribable is that it appears to A that *p*.
c. *The perceptual state is veridical and it is ascribed as non-veridical perception*. It is canonically ascribable that A sees that *p*, but this is not actually ascribed. This is another example of incorrect ascription. For example, it is ascribed that one sees that q. This would happen if the subject had a history of perceptual hallucinations and the ascriber presumes that the subject hallucinates also in the present situation.
d. *The perceptual state is non-veridical and it is ascribed as non-veridical perception*. One has an illusion or a hallucination and while the non-veridical content is canonically ascribable, it is also actually ascribed. This is a correct

ascription of an illusory or hallucinatory content. The main problem with this ascription is to make sure how to determine the non-veridical content. There are no folk-psychological rules for such situations, so presumably one has to descend to the design stance in Dennett's sense. On that level, one can use those regularities that are known about the functioning of the sense organs. There are some known regularities common to all humans – for example, we perceive illusions in the same way, etc.

e. *The perceptual state is a veridical hallucination and it is ascribed as non-veridical perception.* This is like the situation described in (b): A hallucinates that *p*, but it happens that *p* holds as A hallucinates it to be. The difference is that the ascribed state is deemed to be a case of non-veridical perception. 'It appears that *p*' is ascribed. In this case, the ascription may be regarded as canonical. Since we are dealing with a case of hallucination (it does not matter whether it is accidentally veridical or not), then it is right to treat it as such.

f. *The perceptual state is non-veridical and it is ascribed as veridical perception*. It is actually ascribed that A sees that *p*, but it is not the case that *p* and 'A sees that *p*' is not canonically ascribable. This is an incorrect ascription, for there is a mismatch between the content of perception and its object. Instead, 'it appears to A that q' is canonically ascribable.

The singularly compelling source of ascription can be responsible for the situations (b) and (f). For it compels the actual ascriber to start from the environmental object or situation, which leads him to ignore the possibility that the subject might hallucinate.

Note my cautious formulation of these conditions. I do not say that there is a fact that a subject has a veridical perceptual state and that one's ascription misses this fact. For according to the ascription theory, there are no ascription-independent mental facts. Instead, the ascription theorist has to claim that if the fact that *p* and the ascribed content do not coincide, then we do not have a canonical ascription of a factive state. One can say that such and such content was not canonically ascribable. However, this need not always be apparent to actual ascribers. The actual interpretation of minds is thus open to later corrections, where required. From the point of view of the main question of the ascription theory – what is it to have mental state with content – it does not matter whether the content is correct with respect to the world or not. However, since perception is factive, success with respect to the world is written into its individuation conditions and it has to be taken into account in the ascription as well.

One may wonder whether the stress on the object of perception shows that the ascription theory is a species of the causal theory of perception. That theory treats perceptions as caused by external objects. The idea is that the causal element

plays a role in distinguishing veridical perceptual states from non-veridical ones. It is sometimes said that in the case of hallucination and normal perception, there is a common element – the percept or sense-datum – but it is veridical if it is caused by the right kind of object in the environment. An alternative to the causal theory is the disjunctive theory, which does not postulate a common element in normal and hallucinatory perception. Instead, it says that when somebody perceives something, then either it is a veridical perception or it is a non-veridical state, but there is no common state. The disjunctive account is motivated by the rejection of the common element in perception, which one perceives even in the non-veridical case. As a result, an externalist account of perception is presented.

How does the ascription theory relate to this dispute between causal and disjunctivist theories? This is not the place to choose sides. I just want to point out that the approach to perception developed here has features of both accounts. On the one hand, there is some common element in perception; this is protocontent that is available only to the first person. I discuss this notion more closely in the following sections. In addition, recall the importance of the stimulus-source in the ascription of perceptual states. If one wishes, it is possible for one to bring in the causal aspect by saying that the stimulus is causally related to the perceptual state. In terms of the previous chapter, that would be either to make the point that the stimulus is causally efficient in producing the relevant brain processes or that it plays a causally relevant role in the personal-level story about perceptual content. One could also maintain both. On the other hand, factive states and non-veridical perceptual states are not states of the same kind. There are respective differences in the conditions of their ascription as well. Only the ascription of factive states is governed by the factivity constraint, and the ascription of non-veridical states cannot be singularly compelling.

8.4 The content of perceptual states

While the preceding sections presented a general account of what is involved in the ascription of perceptual states and their contents, I will now focus more on the question of what kind of content perceptual content is. This relates to one of the distinctive characteristics of perception, mentioned at the beginning of this chapter: that perceptual states have non-conceptual content, whereas beliefs have conceptual content. In general, non-conceptual content is content that can be ascribed to a state of a creature which does not possess those concepts that are used for specifying the content. This notion can be further clarified with some quotes. For example, Tim Crane (1992: 149) writes that "X is in a state with non-conceptual content iff X does not have to possess the concepts that characterise its

content in order to be in that state." Mark DeBellis (1995) distinguishes between weakly and strongly non-conceptual content. Weakly non-conceptual content does not presume the possession of only those concepts by which this particular content is characterised. Strongly non-conceptual content is the kind of content where the subject "may have an experience as of a certain feature without having any concept at all of that feature" (DeBellis 1995: 70).

Non-conceptual content is introduced because of certain explanatory needs. It is assumed that we require non-conceptual content in order to explain various phenomena (see e.g., Dretske 1981; Evans 1982; Peacocke 1992, 2001; Bermúdez 1994, 1995; Heck 2000). First, there are several considerations from the features of perceptual experience:

a. *The fine-grain of perceptual content*. One can have experiences with contents that are more finely grained than the resources of one's conceptual repertoire. The common examples are the perception of colour shades that one cannot capture by concepts (see Evans 1982: 229), or the discrimination of one silhouette from another without being able to say in what the difference consists (e.g., a range of mountains, see Peacocke 1992: 67–68).
b. *The relative independence of perceptual content from belief content*. There are cases where one's beliefs represent the world as being different from the way it is represented in perception (e.g., the already mentioned Müller-Lyer illusion, see Crane 1992), and despite having such belief, one's perceptions remain unchanged.
c. *The putative common element in the cognition of those creatures that possess concepts and of those that do not.* One may be inclined to ascribe the same perceptual content to creatures that do not possess the relevant concepts as to those creatures that do possess these concepts. By contrast, if perception is construed as an entirely conceptual matter, one has to face the conclusion that animals and infants are creatures that live in a world that is perceptually different from ours.

Second, the considerations in favour of introducing non-conceptual content come from a theory of concepts that is concerned with explicating the possession conditions for observational concepts (Peacocke 1992). Such a theory has to state the possession-conditions of concepts in non-circular terms, that is, without invoking in explanans either covertly or overtly the concepts whose possession is explained, and one way to do this is by appeal to the notion of non-conceptual content.

To consider an example of a theory of perception that involves non-conceptual content, let us review briefly Peacocke's (1992: Chapter 3) account. In his view, an experience may have different types of content, but there is one type of content

that is fundamental in the sense that its existence is needed to give an account of other kinds of content. This content is specified in terms of how the world should be in order for the content of perception to be correct. Peacocke employs the notion of a "spatial type"; this is the type that captures "those ways of filling the space around the subject that are consistent with the correctness of the content." (Peacocke 1992: 62). The space is filled out with properties (surfaces, textures, temperature, brightness, surface orientation, etc.) that are positioned in relation to its origin (for humans, the centre of one's chest) and axes. The concepts that are used to specify the properties by which the space is filled out around the perceiver need not be in the possession of the perceiver. The spatial type, or, as Peacocke dubs it, "scenario", is not composed of concepts, and it is not a proposition; instead, it consists of properties of the external world. In experience, only the spatial type may be given to the subject of experience, not the concepts by which that type is described (see also Peacocke 1994: 420). That is why scenario-content counts as a species of *non-conceptual* content.

Peacocke has employed his notion of scenario-content to explain the issues listed in the previous section. In addition, he uses this notion to explain concept possession. According to him, the nature of a concept consists in the capacity to master that concept, as explicated by the "possession conditions" for that concept. The possession conditions are given in terms of a subject's tendency to find certain kinds of inferences containing the concept compelling, and, in order to avoid circularity, the conditions have to be specified without containing in the range of the subject's propositional attitudes the very concept for which possession conditions are given (Peacocke 1992: 71). Non-conceptual content becomes important in giving the possession conditions for perceptual concepts. For example, the possession condition for a perceptual concept *C* says that "it is that concept *C* to possess which a thinker must be willing to judge that certain things are *C* in such and such circumstances in which he perceptually experiences them as falling under *C*, etc." (Peacocke 1992: 89). The specification of what is involved in "falling under the concept" mentions non-conceptual content and it can thereby avoid circularity. The circularity is avoided since the specification of non-conceptual content does not require that the subject possesses those concepts by which it is specified.

There are thus various reasons for assuming that perceptual states, both factive and non-factive, have non-conceptual content. Of course, the issue of whether such reasons are indeed conclusive is currently hotly debated. It is beyond the theme of the present book to review this discussion. Instead, I inquire about how the ascription theory can accommodate non-conceptual content, given the assumption that perceptions have non-conceptual contents. I think that there are no principled reasons why an interpretivist could not allow non-conceptual contents. An alternative would be to analyse perception solely in terms of perceptual

beliefs, but this is problematic precisely because of the aforementioned differences between belief states and perceptions. Hence, it makes sense to consider the alternative approach in terms of non-conceptual content, and to this purpose the next section is devoted.

8.5 Non-conceptual content in the ascription theory

The main impetus for making sense of non-conceptual content in the ascription theory comes from the previously mentioned set of issues and not from its place in explaining concept possession. The ascription theory does not have to introduce non-conceptual content in order to avoid circularity in explaining concept possession, since it does not individuate concepts in terms of their possession. In addition, the ascription theory cannot accept a certain reading of the Peacocke's account of concepts. Namely, if contents are composed of concepts and concepts are individuated through their possession-conditions in an ascription-independent manner, then that would generate a conflict with the ascriptionist account of content possession. In that case, there would be an ascription-independent account of what is involved in the possession of content and it would be explained in terms of making certain inferences and having certain perceptions.

Instead, the ascription theory takes concept possession as derivative from the ascription of propositional attitudes.[3] On this construal, some of Peacocke's possession conditions can also be embraced by the ascription theory, for they can be viewed as explicating the common-sense connections between various abilities and behavioural manifestations that are involved in a concept. If the respective behaviour occurs and such abilities become manifest, then the possession of a given concept is ascribable. In this account, concept possession is not a deep fact about a person. It is rather an abbreviation of a longer story about what is ascribable to someone. For example, if we say that someone possesses the concept 'red', then this just abbreviates claims about the ascribability of the contents containing the concept 'red'. When these latter claims are unpacked in turn, we learn, for example, that this person can distinguish red surfaces from green surfaces, masters some inferences concerning this concept and so on. It follows from this that concept possession cannot be used as a prior and independent constraint on the possession of certain mental phenomena, for both types

3. This presumes a distinction between concept possession as a psychological matter, on the one hand, and the meaning of words as a semantic matter, on the other. As the ascription theory exlains the possession of mental states, it can explain concept *possession*; it does not explain the meanings of words.

of possession depend on the same kind of ascription conditions. This idea comes close to Alex Barber's deflationary approach to concept possession, which has the following consequence:

> If we have a defeasible reason for thinking that an agent believes some proposition, then this reason cannot be defeated solely by the claim that the agent does not possess some concept involved in the allegedly believed proposition, because any reason for belief attribution is *ipso facto* a reason for concept possession ascription. (Barber 1998: 57)

However, a good case for non-conceptual content can also be made in interpretivist terms. Some notion of non-conceptual content is needed when there is a reason to ascribe content to a subject, but the concepts by which this content is characterised are not ascribable to the subject. This is similar to the reason (c) above. There are cases – for example, situations involving higher animals or humans in pre-linguistic stages of development – where the subject lacks sufficient cognitive sophistication to warrant the ascription of concept possession, but where it still makes sense to ascribe some contents to the subject. The notion of non-conceptual content would allow us to say that there is a common element in the perceptual states of normal humans and non-verbal subjects. A simple example is when we ascribe to a cat that it sees that there is a portion of Whiskas on its plate. The perceptual state with a similarly specified content can also be ascribed to someone who is feeding the cat. In the case of the cat seeing Whiskas, we ascribe contents even when the concepts by which these contents are specified are not ascribable. There are also other cases in which the concepts used to specify perceptual contents are available only to a sophisticated theorist. In these situations, normal people are on a par with non-verbal animals. Such is the case when a theorist makes use of a subpersonal account to explain one's perceptual functioning, and ascribes contents to the subject and not only to its subpersonal parts.[4] In such a situation, a state with content is ascribed to one that is specified by concepts whose possession becomes ascribable to the person only after the person has learned the theory in which these concepts are embedded.

The distinction between non-conceptual and conceptual content is often drawn with the help of Gareth Evans's Generality Constraint. This is a general constraint on the recombinability of the conceptual elements of thought. The ability of thinking structured thoughts is general in the sense that once one has the

4. In addition to such cases, there is a type of subpersonal explanation that involves ascribing contents to certain stages of subpersonal processing. For example, one can say that such and such subpersonal states compute a certain kind of information. Some authors (e.g., Bermúdez 1995) take this to be an ascription of non-conceptual content as well.

ability to entertain all the structural parts of the content, one is capable of entertaining all the combinations of those parts: "If a subject can be credited with the thought that *a* is *F*, then he must have the conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception." (Evans 1982: 104.) Drawing on this constraint on conceptual thought, we may say that non-conceptual thought does not exhibit such systematicity. When we interpret a subject by ascribing non-conceptual content to him, this does not licence any further content attributions, which would be licenced if the subject's thinking were governed by the Generality Constraint. For example, when one ascribes the non-conceptual thought to a cat that (described in our terms) there is food on the plate and that the plate is in the kitchen, it does not follow that the non-conceptual thought that the food is in the kitchen is ascribable to that cat. In the case of the ascription of thought with conceptual content, the content that violates the Generality Constraint is normally not ascribable (unless there are special overriding reasons, as in the case of some aphasias, for example).

Since the central claim of the ascription theory is that the possession of mental states and contents ineliminably involves ascribability, the ascription theory should deny the ascription-independent possession of non-conceptual content. I shall call the ascribable non-conceptual content '*shallow* non-conceptual content'. The type of content whose possession is not constitutively connected with ascription may be called '*deep* non-conceptual content'. The fact that one's states have deep non-conceptual content is viewed as a natural fact that occurs independently of any interpretation. The ascription theorist has to deny deep non-conceptual content. The problem with such deep content is that non-conceptual content is specifiable and this conflicts with the "deepness" of this content. As the notion of non-conceptual content is understood, each token of non-conceptual content has a specification in some terms, although the possession of the respective concepts is not ascribable to the subject. The specification gives the correctness conditions for the content. However, specifiability brings along ascribability. The minimal requirement for non-conceptual content to be ascribable is that there must be a way of picking out the content of perception. This can be done either by using demonstratives or some more coarse-grained terms.[5] For example, a certain perceptual

5. This assumption may seem to go against the initial motivation for the very notion of non-conceptual content: to explain the fineness of grain of perception. On my construal, although shallow non-conceptual content can be used to explain the fine-grained representational aspects of perception, it does not capture the phenomenology that can accompany the content. Sometimes these two aspects are not distinguished in relation to the fineness of the perceptual grain. However, my view is that the fine-grain of perception has phenomenal aspects that are not representational, and I use a different construct to explain this.

content can be ascribed by saying in the presence of a colour sample that X sees this rose as being that shade of red. This is not to refer to the phenomenal character of the experience; rather, it is to assign non-conceptual demonstrative content to the perceptual state. As before, the demonstrative content is non-conceptual only if the possession of the required demonstrative concepts is not ascribable to the subject. If it is ascribable, the content of the perceptual state, specified in the demonstrative manner, is conceptual.

Deep non-conceptual content, on the other hand, is inherently representational and bears no specification essentially, that is, no specification would individuate what is represented. For if there were such a specification, it would be usable for the ascription. However, deep non-conceptual content is supposed to be such that it cannot be fully specified by concepts and it is difficult to see how this can be the case. There are no corresponding problems with shallow non-conceptual content, which is dependent on the ascription that specifies in theoretical or other terms what is represented by that content to the subject to whom it is represented in the non-conceptual manner.

The motivation for thinking that non-conceptual content is deep comes in part from introspective or phenomenal consciousness. The notion of non-conceptual content is mixed with what is introspectively given to one, and that surely seems to be independent from anything that an interpreter might say. But the representational and the introspectible aspects should be kept apart.[6] Non-conceptual content is a theoretical notion that captures the idea that one can have representational powers without the possession of concepts. (Or, in interpretivist terms, the representation of the aspects of the world can be ascribable to one, while the possession of the concepts that specify this representation is not ascribable.) On the other hand, the notion of what is present in experience is not a theoretical notion, and there is no reason to assume that the former notion has to explain the latter. For the experientially present, I propose to use another explanatory construct, the notion of *protocontent*. This is what is given to one in consciousness and it is distinguished from the content proper for philosophical purposes. One has protocontent irrespective of ascribability. In what follows, I attempt to explicate this notion further.

6. Cf. this with a related distinction by Crane (1992): one notion of non-conceptual is given in terms of the *possession* of concepts and the other is given in terms of the *constituents* of content. On this other reading, which Crane rejects, content is non-conceptual if it is not composed of concepts. See also Maund (2003: 158–159), who distinguishes content that does not require the possession or application of concepts from the "non-conceptual component or ingredient of content". I do not take the 'component' notion of the non-conceptual to be a representational notion.

8.6 Phenomenology, non-conceptual content and protocontent

The notion of content is an ambiguous notion. In one sense, it expresses what is represented. In the second sense, it may be taken to mean what is consciously presented to the subject – the given experiential world. I assume that only in the first sense may we legitimately talk about content and that it is better to distinguish this sense sharply from its second, presentational sense. The reason for the distinction is that the content that is represented need not be consciously presented and not all that is presented in consciousness need represent. As for the first part, we can make sense of and ascribe representations that have no connection with conscious presentation, e.g., unconscious representations and nonconscious representations like sentences, signs, etc. As for the second, some aspects of conscious experience might be nonrepresentational. This is more controversial, though, but the very existence of a controversy shows that there must at least be a conceptual distinction. Otherwise, the issue would reduce to a mere matter of grasping the notion of conscious experience. In addition, it is unclear in what sense that which is presented to a subject can be a representation. Usually, we say that something is presented when it is given to one, like a flower or a book. But this is not the representational notion. Perhaps it is implied that the subject has access to that which is presented to him, and the idea is that there are some internal images which the subject can access and which themselves represent entities in the external world. But the picture theory of representation is deeply problematic (see e.g., Cummins 1989). In particular, there is no reason to assume that in order to have something represented to the subject that it has to be mediated by some internal image.

Protocontent is content in the presentational sense of the term: it is not a representational construct.[7] It is defined in this way. Above all, the possession of protocontent is not the possession of content in the representational sense. To have protocontent is to have certain subpersonal processes going on in one's brain and it is characteristic of these processes that there is something like having them. In one respect, protocontent is not a mental occurrence, especially if we restrict the notion of 'mental' to those occurrences and dispositions that are individuated at the personal level and in folk-psychological terms. For each attempt to describe protocontent in mental terms would only give yet another ascription of a mental state or content as a result, but, by definition, protocontent is something that is not exhausted by any such description. In another respect, it seems that it deserves to be called mental, for it involves phenomenality. Yet again, it does not

7. Thus, it is to be distinguished from Peacocke's (1992: 77) "protopropositional content", which is a kind of non-conceptual representational content that links individuals with properties or relations.

have any straightforward folk-psychological specification, for it is a philosophical idealisation, a result of a certain conceptual dissection performed on the richer notion of experience. In this respect, protocontent is not quite the same as the total, phenomenally given content of consciousness, for the latter is presumably a mixture of phenomenal experience and self-ascriptions. Protocontent is rather phenomenal experience minus the judgement that takes it to be of one mental kind or another. We can distinguish between first-order judgements and reflective judgements. In the first case, the content of a judgement does not involve the state under which this content is apprehended. For example, this is to judge that something is red. In the second case, the content of a judgement involves some mention of the state: this is to judge that I see something to be red. Not only is protocontent experience without reflective judgement, it is also experience without first-order judgement. Such a separation can be made in theory; it is not so clear whether one can do this in one's awareness. I assume that it is a matter of dispute whether one can be aware of protocontent barely, such that awareness involves no self-ascription of content or a state.[8] But one would be aware of protocontent by being aware of one's experience.

The claim that protocontent is what it is like to undergo certain brain processes is in accord with my naturalistic inclinations. It does not entail the dualism of inflationary properties. Given that protocontent cannot be brought under any mental predicate, it is not a mental property. If, on the other hand, it can be picked out in mental terms, this only commits us to pleonastic mental properties to which the familiar interpretivist story applies. There is no reason to assume that inflationary mental properties are implicated here, for there is no direct route from phenomenal appearing to the existence of an inflationary property. For example, in the inflationary sense of the term, it does not follow from the fact that it feels a certain way to have experiences that experience has the inflationary properties of qualia. Having protocontent is having certain natural properties.

Views tending in the same direction have been expressed by several authors. For example, Putnam (1981: 81) proposes that a functionalist should say that "the 'qualitative character' is just the physical realization." Heil (2003a: 235) notes that "the qualities of conscious experiences are perfectly ordinary qualities of brains." Note also that since having protocontent is having objective brain processes, the

8. Dennett opposes the idea that one can distinguish between how it seems and how it is judged to be seeming. I agree that this distinction may not make much sense introspectively, if one can access one's experiences only through judgements. However, it does not follow that we cannot make such a distinction in a theory. For a view that distinguishes access and phenomenology, see Block (1995).

possession of protocontent is independent from ascribability, just as one's possession of brain processes is independent from interpretation.

The details of these subpersonal processes need not concern us here. I want to point out only that the notion of protocontent need not commit one to a kind of materialist Cartesian picture of mind, criticised persuasively by Dennett (1991).[9] It does not depend on notions similar to the Cartesian Theatre or the Cartesian Subject. Less metaphorically, saying that one has protocontent does not presume that there must be a subject perceiving it. Neither does it presume that there is a special spatial location in brain where processes become conscious. Finally, protocontent has no nonarbitrary beginning and end in time: it does not make sense to individuate protocontent temporally, to point to a specific moment when one protocontent ends and another begins. 'Protocontent' has no plural in respect of a single subject.

Owen Flanagan (1992: 58) has said that "some patterns of neural activity result in phenomenological experience; other patterns do not. The story bottoms out there." I agree with the first part, but I do not think that the production of phenomenal experience in the brain should be taken as a primitive. Protocontent can be further decomposed by functional or subpersonal analysis. However, there is no reason to assume that some functional or neural unit should correspond to protocontent. Protocontent is not a special organ in the brain. Experience has a characteristic phenomenal unity, but it may well turn out that this unity is composed of distinct microepisodes. From the philosophical viewpoint, we may take protocontent as a black box (or a phenomenal box, to be exact); a placeholder to be filled with an account in neural terms. This is permissible, in the present setting, insofar as the ascription theory does not purport to be an account of consciousness or an account of the processes whereby mental states become conscious.

The reason for introducing this notion into the ascription theory is to make the theory phenomenologically more adequate. It is not an *ad hoc* feature, though, for its introduction is to be expected once one reflects on the dialectical situation of the ascription theory. This account views mind as an interpretational construct, constituted by the common-sense framework that guides the application of mental terms. Mental discourse forms a more or less intertwined conceptual system that can be applied to suitable subjects. If mind is understood in this way, then the scientific value of mental constructs is limited to the scientific value of folk-psychology. The ascription theory, which is a philosophical meta-theory of

9. I shall not discuss Dennett's Multiple Drafts model of consciousness separately here, for it is a subpersonal account, and thus removed from my more immediate aims. Note that contrary to a widespread assumption, Dennett is not an eliminativist about experience (see the remark in Dennett 1995: 289).

folk-psychology, does not assume that its constructs should be used in scientific psychology in order to explain the causally efficient processes that take place in the subjects of ascription. This does not diminish the everyday explanatory relevance of mental constructs. But if such a distinction is made between the ascribed mind and its causally efficient basis, then what remains is experience as it is given to the subject himself. On the one hand, pure ascribability fails to do justice to the experiential viewpoint, for experience has aspects that are not ascription-dependent. On the other hand, we still lack an understanding of how experience originates from processes in the brain. In this situation, protocontent serves as something that is left uncaptured by the application of mental terms, but at the same time, by maintaining that it is a special sort of brain process, we can proceed to give a naturalistic account of experience.

Experience has phenomenal features. It is a matter of lively debate how to understand these features exactly, but the radical claim that experience has no phenomenality is too counterintuitive to take it seriously. Sometimes the issue is set up as a battle between two camps. In the one camp, there are those who claim that experience either has only representational properties or that all phenomenal properties of experience supervene on and are explained by its representational features (see e.g., Dretske 1995). Others say that experience also has sensational properties that cannot be reduced to representational properties (see e.g., Block 2003b). This issue seems crucial in battles over physicalism only if the inflationary conception of a property is presumed. Since I presume that those inflationary properties that belong to the natural stratum include neither representational nor sensational properties, I shall not take a stand in this dispute.[10] Instead, I will try to explain how phenomenal features can flow from the nature of protocontent, without assuming that they are either sensational or representational inflationary properties.

We can view these phenomenal features as the properties of protocontent. As protocontent is a brain process, the properties of protocontent are *ipso facto* properties of brain processes. Speaking metaphysically, the natural basis of protocontent is a causal power in the brain, but it can be described in various ways, giving rise to different deflationary properties. For example, we could say that protocontent is ineffable, and we thereby ascribe a pleonastic property to its natural basis. Nevertheless, we should come up with an account of what in the natural basis makes it the case that we can make such ascriptions. The idea is that if one were

10. However, I submit that to say that experience has phenomenal features which require explanation is not yet to say that there are qualia, if we take 'qualia' to denote inflationary properties that are intrinsic, ineffable, private and directly consciously accessible. Such properties are philosophical inventions and should be quined (see Dennett 1988).

to have explained why protocontent has the properties it has, one would have also explained why phenomenal features of experience are as they are. The relevant properties of protocontent that I have in mind are *continuousness* and *ineffability*. Protocontent is continuous in the sense that it cannot be captured entirely by such discrete entities as words. For this reason, it is also ineffable and thus cannot be communicated to others. All we can communicate is our self-ascriptions, but there is an intuition that something always remains unsaid. I assume that these two features of experience have done much to mystify the notion of qualia. For many, it seems difficult to accept that there are some facts or properties that are ineffable and continuous. However, at the same time, the introspective pondering of experience seems to lead to such a conclusion. But if we view these properties as belonging to the format of a mode of presentation of certain subpersonal processes, then much of the troublesome character of qualia seems to fade away. Viewing ineffability and continuousness as arising from the format of subpersonal processing, we can make sense of them from the naturalist point of view. Of course, what remains to be explained is why certain processes in the brain give rise to such a presentational mode and this is a substantial task on its own.[11]

Protocontent is (normally) only part of the experience of the person himself, since (normally) no one can undergo someone else's brain processes. (The normality clause is added since one day it may become feasible to hook up two brains to each other so that these brains can undergo each other's processes.) Protocontent is thus an issue only in the case of self-ascription. An account of how protocontent can be implicated in self-ascription is provided in Chapter 9. Here I want to discuss the relation of protocontent to conceptual and non-conceptual content. Let us begin with the differences between third- and first-person ascription of perceptual content. One difference should already be apparent: the first-person's protocontent is not accessible from the third-person point of view. However, this has no effect on the overall structure of the ascription theory, since protocontent is not among the sources of ascription: it does not provide reasons for the ascription. We will return to this point later.

Another difference is that one can self-ascribe only conceptual contents to oneself. This follows from the notion of non-conceptual content, for one does not possess the concepts that specify this content, and as one does not possess those concepts, one cannot self-ascribe the content to oneself that is specified by these concepts. Indeed, the failure of self-ascription is one ground for the ascription of

11. Cf. Jakab (2000) for an attempt to explain the ineffability of phenomenal features as a result of certain representational and computational mechanisms in the brain. Although I do not agree with the details of his account (for a criticism of Jakab, see Metzinger and Walde 2000), it is a good example of the approach to phenomenal features that I have suggested here.

the lack of possession of the concept. Recall that concept-possession talk is an abbreviation of a longer story about the subject. This longer story would also involve the claim that there are terms that express given concepts, and that the subject is able to apply these terms in particular circumstances. The ability to use mental terms to describe one's experiences is just one component of the possession of concepts expressed by that term. That the term is a particular should not lead one to think that the possession of a concept is the possession of something particular.

This relates to Dretske's (1995) claim that one needs concepts in order to be aware of one's experiences as experiences.[12] Dretske (1993) distinguishes between seeing and seeing as, or the awareness of *x* from the awareness that *x*. The latter requires that the concept of *x* is available to one's thought. There can thus be experiences inaccessible to a subject: they are those for which the subject lacks concepts. In the terms of the ascription theory, this is to say that one cannot self-ascribe experiences with non-conceptual content if one does not possess the required concepts (that is, if the possession of such concepts is not ascribable to one). However, this non-conceptual content is *ascribable* to this subject from the third-person perspective, by an ascriber who possesses these concepts. Non-conceptual content is ascribable to a person, even if the person cannot ascribe it to himself. The first-person has protocontent, of course, but this is not sufficient for having content, either conceptual or non-conceptual in the shallow sense.

It also follows from the fact that one cannot self-ascribe non-conceptual content that this content cannot be the subject's own reason for a belief. This is the main reason why Bill Brewer (1999: 150–152) rejects non-conceptual content and claims that perception must be conceptual. He argues that reason-giving requires propositional premises and conclusions and that in order for a reason to be a subject's own reason, this proposition has to make up the content of the subject's mental state. This requirement is not satisfied in the case of non-conceptual content, since, he argues, the content of one's mental states is not propositional. I think that Brewer's criticism holds good against deep non-conceptual content. As concerns shallow non-conceptual content, it is true that it is not a subject's reason from her own perspective, but this does not mean that it cannot be used as a reason in the ascription made by the third person. The third-person ascriber can apply the common-sense justificatory relation to the subject by saying that since a certain non-conceptual content is ascribable to one, a certain perceptual belief is also ascribable to one. The ascriber can even say that the subject can be viewed as holding this belief for the reason that she has this non-conceptual content. There

12. Dretske (1995: 156) writes, "You do not need the concept RED to see red, to experience this quality. But you *do* need this concept to become aware of the *quale* red, to become aware that you are having an experience of this sort."

is no requirement that the reasons be conceptualised by the subject herself (see also Hurley 2001: 425).

Let us come back to protocontent. It may certainly seem puzzling as to why protocontent itself cannot be a reason for ascription. Moreover, one might be tempted to hold the following reason-giving view about protocontent: having protocontent is in the first-person case a reason for self-ascription, and such content is ascribed that corresponds to that protocontent. In a sense, one obstacle to this view was already removed by the introduction of the notion of the singularly compelling ascription. Protocontent would be, it seems, a singularly compelling source for the ascription.

However, I reject this view, since I cannot make sense of protocontent as a source for ascription. I will elaborate. If protocontent is taken as a source or a reason for an ascription, then we would have two options. Let us call these options 'theory A' and 'theory B'.

In theory A, protocontent is the undifferentiated Given. It is like the "cosmic porridge" described by Kirk (1999: 52): there exists an "indeterminate something" that acquires determinate properties through one's acts of conceptualisation and classification. In this theory, protocontent is turned into content when one applies concepts to it during the act of ascription. But at the same time, the Given is a reason for ascription, for if protocontent would have been different, different content would have been constructed. However, this theory has obvious faults. First, there is no account as to what guides the ascription when protocontent is undifferentiated porridge, for it does not compel one to make any specific ascription. Second, it is unclear how something that is undifferentiated could provide reasons. This would be a version of the view attacked by Brewer (1999) and McDowell (1994). While in the case of non-conceptual content, there was an answer to this criticism, namely that the non-conceptual content has a conceptual specification provided by the third-person, on this understanding of protocontent, no such specification is even forthcoming.

By contrast, theory B takes seriously the assumption that protocontent compels a person to self-ascribe one content and not another. As a result, this theory says that protocontent should itself be already conceptually determined and then guide the ascription by its conceptual features. In this view, protocontent would already be a kind of representational content. But if protocontent is already conceptual, then we can say that to have protocontent is simply to have content and there is no need to construct this in terms of the ascription theory. Theory B of protocontent is thus potentially damaging to the ascriptionist approach as a whole. Of course, one way for the ascription theorist to counter the claim that the ascription is not needed is to say that the ascription personalises the content, that it adds to the content that it belongs to a certain person. For example, it turns

the content 'This is a nice rose' into the ascribed state 'This is a nice rose and A believes this'. However, I think that such a manoeuvre would make the ascriptionist account too weak, leaving it only the personalising role, so I think that a better answer for an ascription theorist is to reject the reason-giving view itself.

My view is that protocontent is not evidentially related to ascription and it does not guide the ascription in a conceptual or justificatory manner. In my view, one does not have to apply concepts to sensuous material (whatever that might be) in order to have conceptual contents. Contents are not constructions out of protocontent. One has contents whenever they are canonically ascribable to one. Protocontent neither precludes having conceptual and non-conceptual contents nor is it a precondition for the possession of content. In particular, it does not make sense to try to specify the content of protocontent, for there is no measure of correctness here. Similarly, it makes no sense to say that one's specification of protocontent was incorrect. The mistake made by such claims is the presumption that our specifications can fix the content of protocontent, but protocontent is not something that can be turned into content by providing its specification.[13] Alternatively, one might say that when one makes an ascription at time *t*, this ascribes content to the protocontent one has at time *t*, but this makes little sense. Otherwise, we would have to say that some ascribed contents express protocontent better than others, and this would confer a reason-giving role on protocontent.

I thus reject both theory A and theory B. Theory A, which claims that the undifferentiated Given provides reasons for differentiated contents is just hard to take seriously, for there is nothing to determine the content of a reason in such a case. One cannot say what the reason for ascribing content *p* rather that ascribing content *q* is if one proceeds from theory A, since the reason itself is an undifferentiated "porridge". For reasons brought out in the literature against the notion of the Given and the scheme-content distinction, we should avoid committing ourselves to the "dual capacity" account that pictures the perception as a blend of the sensuous material and the concepts applied to it (see the discussion of Sellars (1956) in Sedivy 2004, as well as Davidson 1973–74). The main problem with such an account is that, on the one hand, the sensuous material is supposed to have already an intrinsic character that guides its classification in a certain way, thus being already somehow conceptually structured, and, on the other hand, the conceptual classification should be external to the sensuous material, which

13. Note that content specification is a different matter to describing the features of experiences, which allow us to make claims about the form of protocontent (e.g., to say that it is ineffable). The latter is a claim about protocontent as a presentational notion and not about the representational construct.

must be the unconceptualised causal effect of the world on us. Arguably these two conditions cannot be jointly satisfied, for nothing can be both conceptually structured and unstructured.

John McDowell (1994), drawing on Sellars (1956), has argued against non-conceptual content on the ground that it is a version of the Myth of the Given. He assumes that the Sellarsian attack on the Myth of the Given established that there is no given element in perception. I have already pointed out that shallow non-conceptual content is not guilty of committing one to the Myth of the Given. Another question is whether my notion of protocontent is vulnerable to this attack. If we understood protocontent as theory A suggests, then, as already noted, the criticism of the Given would apply to it. But this understanding of protocontent should be rejected. Note also that Sellars attacked those foundationalist and logical atomist aspects that were traditionally connected with the given element, but which are not be shared by the ascription theory. For example, it was assumed, especially in the sense-data theory, that perception presents one with some internally given items that are non-inferentially known and that provide the warranted basis for all empirical knowledge. McDowell's and Sellars' arguments dispute this idea (for a close and critical analysis of this, see Bonevac 2002). However, nothing of the kind is presupposed here. Protocontent is not a part of the justification of empirical knowledge.

As concerns theory B, then, I reject the account since, as the notion was introduced, protocontent cannot be conceptual in the manner required by theory B. Moreover, if we stick to the view on concept possession that was outlined earlier in this section, the issue of whether protocontent is conceptual or not loses its force. For given that the possession of concepts is constituted by the same principles that constitute the ascription of contents, there is no ascription-independent question of whether one has a given concept or not. Similarly, there is no ascription-independent question of whether protocontent is conceptual or not.

There are two ways in which one can make sense of protocontent in theory B. The first option is to understand it as an ordinary contentful mental state, and, as stated, this approach risks rendering the whole ascription theory obsolete. As the aim is to make sense of protocontent while presuming the ascriptionist approach, this option is not feasible. The second option is to take protocontent as non-conceptual content. Since having protocontent is independent from ascribability, the non-conceptual content must be deep in the sense outlined above. However, the notion of deep non-conceptual content is as confused as the notion of the undifferentiated given that has a conceptual structure (cf. Sedivy 2004: 172). Moreover, since the postulation of deep non-conceptual content conflicts with the ascription theory, one cannot use it in formulating an ascriptionist account of experience. In my view, there are no explanatory tasks

for deep non-conceptual content that cannot be fulfilled by the notions of shallow non-conceptual content and protocontent.

To summarise: in this chapter, I attempted to extend the ascriptionist approach to perceptual states and experiences. The examination of the differences between beliefs and perceptions led to the discussion of the distinctive conditions of the ascription of perceptual states. I also pointed out that the ascription theory can accommodate the notion of non-conceptual content when it is suitably re-interpreted. Finally, the notion of protocontent was introduced and discussed with the aim of doing justice to common-sense intuitions about the phenomenal aspects of experience. My contention is that such intuitions can be respected and an interpretivist need not deny phenomenal experiences or remain silent about them.

CHAPTER 9

Self-knowledge

Self-knowledge is a touchstone for any interpretative account, for it is here that one finds the strongest clash with ordinary intuitions. There is a presumption that each of us has immediate and authoritative access to their own mental states. This presumption seems to conflict with the interpretivist approach which has a distinctive third-personal feel, for if the interpreter is in a position to determine which mental states one has, then presumably this leaves no epistemic privilege to the first person.

My main purpose in this chapter is to argue that the interpretivist framework can accommodate and explain the special features of self-knowledge. I will do this by presenting an ascriptionist account of one's access to one's mental states and contents. The point is not that the interpretivist should just say something about self-knowledge. What the interpretivist says should be adequate and acceptable. Richard Moran (1994: 156) has outlined two adequacy conditions for an account of self-knowledge. First, an adequate theory should do justice to the intuitive difference between the first- and third-person case. This difference is captured by the observation that the first person is authoritative in respect of their own mind. Note that this is not yet to say that one is infallible, incorrigible or in some way omniscient about one's own mind. The task is thus to explain what constitutes this authority. The second condition is that this explanation should proceed without entailing a revision in our concepts. The idea is that there are no specific first-person meanings for mental terms; the terms have the same meaning when applied to oneself and to others. This requirement can be met if we presume that both self-ascriptions and other-ascriptions make use of common mental terms. This assumption becomes obligatory if the aim is to have a single theory for both first-person and third-person ascriptions. Further, the general ascription-guiding principles and sources should also be the same in the first-person and the third-person case. If these differed, there would arise familiar sceptical problems with other minds. The overall aim of the chapter is thus to provide an account of the authority of self-knowledge that does not compromise the overall structure of the theory.

Another issue that presents problems for an interpretivist is the role of subpersonal mechanisms in the production of self-knowledge. Does an interpretative account rule out any subpersonal explanation of self-knowledge and, if it does not, then can it be reconciled with some subpersonal explanation? *Prima facie*, this

seems difficult if not impossible, for, on pain of overdetermination, the objective subpersonal account should thus be regarded as the whole story about self-knowledge. In other words, if we already have a subpersonal account of self-knowledge, then any additional description of how one comes to have knowledge about one's own mind should stem from the subpersonal specification. The interpretivist is thus challenged to provide a theoretical role for subpersonal processes in the production of self-knowledge that does not rule out the interpretive account.

The plan of the chapter is the following. I will begin with some stage-setting, presenting a short overview of the accounts of self-knowledge. Then I discuss Crispin Wright's approach more closely and this discussion leads naturally to the aforementioned aim of providing an ascriptionist account of self-knowledge that leaves room for both subpersonal processes and first-person authority. As in the case of perceptual content, I can only present the account in broad outline in this book. Its thoroughgoing defence, comparison with competing approaches and additional developments remain for another occasion.

9.1 Approaches to self-knowledge

I begin with some general remarks about present discussions concerning self-knowledge. It has to be noted that most discussions of self-knowledge, although there are exceptions, do not spend much time on the epistemic question of what makes self-knowledge *knowledge* as opposed to belief about one's mental states. Instead, several models are proposed for how one arrives at beliefs about one's mental states and whether these beliefs are a priori or not. The higher standards involved in the notion of knowledge are not attended to. I will use the notion of self-knowledge in the similarly minimal sense that there are truths about one's own mental states that one can know and, apart from a short discussion in the end, I shall not pay much attention to the notions of truth and knowledge involved.

There are two basic ways to treat self-knowledge. Accounts of self-knowledge can be either cognitively substantial or insubstantial (Boghossian 1989). In the cognitively substantial approach, self-knowledge is a matter of *self-observation* or an *inferential* process based on internal evidence or inner reasons. For example, if we treat self-knowledge in the perceptual model, we have to assume the existence of inner and private mental events that are apprehended by some internal scanner. In some circles, this view is despised as a version of Cartesianism. Usually, however, Cartesianism is taken as the stronger view that postulates inner and private counterparts of material objects and events (like the mental event of seeming red, the inner episode of believing that it is raining, the mental image of a table, and so on) and self-knowledge is construed as the infallible acquaintance with such

inner and private objects. There are decisive arguments to show that such an account is not tenable and, in fact, most contemporary accounts of self-knowledge can be viewed as reactions to Cartesianism. Both the inferential account of self-knowledge and the self-observation view are substantial, since the agent has to do something in order to earn the claim to self-knowledge. However, it is common contention that such views have too many problems to be credible (see e.g., the criticism in Boghossian 1989).

The alternative, then, is to deny that self-knowledge is a cognitive achievement, accomplished by drawing inferences, or to deny that first-person judgements must be based on introspectively accessible evidence. This is to view self-knowledge as cognitively insubstantial. There are various ways in which such an approach can proceed. For example, one can just take the other side of the Cartesian coin and exteriorise the mind in a *behaviourist* manner. This makes first-person epistemology insubstantial indeed, but it is not a sensible position either. The claim that mental state reports are made true by observable behaviour cannot survive critique. It leaves us with no good explanation of the authority and the seeming effortlessness of first-person judgements about one's own mind. Another option is to claim that self-knowledge is not based on reasons. This is to adopt a reliabilist view of knowledge: on this view, our beliefs about our own mental states qualify as knowledge, since they are the results of some reliable processes. I shall discuss the reliabilist approach in Section 9.4.

There is a yet further option in the insubstantialist camp: *non-cognitivism*, a view, which is sometimes attributed to Wittgenstein, that expressions of mental states or avowals are authoritative because of their "grammar". It is claimed to belong to the nature of first-person utterances about mental states that they are authoritative about their subject matter. The reason for this, it is said, is that such avowals are part of the very mental states they express. Just as blushing is part of the condition of being embarrassed, so are non-inferential first-person expressions part of the expressed mental states. This non-cognitivist approach is known as *expressivism*, stressing the idea that avowals are not reports of expression-independent facts about mental states. This approach treats avowals like performances and not as assertions: avowals are not propositional and they lack truth conditions. In this respect, self-knowledge is not knowledge; it is rather a non-cognitive ability to make avowals.

Another approach to self-knowledge that does not take it as a substantial achievement is the *self-constitutivist* account advocated by Crispin Wright (1989a, 1989b). He presents this as an elaboration of the Wittgensteinian insight that we should reject the picture in which the truth of a statement *always* consists in independent facts corresponding to the statement (Wright 1998: 117). Instead, the truth for statements about one's own mind is self-constituted. The key to

understanding the possibility of self-knowledge is thus to take self-ascription as constitutive of the truth of the ascription. Similarly to non-cognitivism, this approach also stresses the idea that the authority of self-ascriptions is a conceptual feature of our mental vocabulary, but it does not assume that beliefs about one's mental states are not truth-apt. In the following section, I discuss Wright's account in more detail, for it provides us with some elements that are incorporated into my ascriptionist theory in the second part of this chapter.

9.2 Self-constitution

In order to understand Wright's proposal about self-knowledge, we have to explicate some of his background assumptions. I present his basic ideas only briefly, omitting from the discussion further criticism. After all, the aim is to set the stage for the ensuing ascriptionist treatment of self-knowledge.

Wright (1987, 1989b) notes that for certain classes of concepts, the truth-values of statements involving these concepts are determined by our best judgements about them. Wright draws a distinction between *extension-determining* and *extension-tracking* judgements. In the extension-determining case, our best judgements or beliefs about a given subject-matter are true judgements: they determine what counts as true in this area, or as Wright (1989b: 248) says, they "determine the extension of the truth-predicate". Extension-tracking judgements, on the other hand, merely reflect the extension of the truth-predicate, which is constituted independently of facts about our best opinions. Hence, they can be the best we can achieve and yet not be true. Wright (1993: 77) expresses the extension-determination idea with the help of *basic* equations in which p is a proposition or a statement involving some concept under study:

$$p \leftrightarrow \forall X\ (C \rightarrow X \text{ believes that } p)$$

According to Wright, our beliefs determine the truth-value of statements involving certain concepts only if this equation expresses an a priori truth. '*C*' points to certain conditions that have to obtain in order for X to make the required judgement (cf. Wright 1989b: 247–248). These conditions have to be specified non-trivially and facts about their satisfaction should be logically independent from facts about the extension of the concept, for which the basic equation is formulated.[1] Wright's

1. Otherwise, the equation would fail to be extension-determining. For details, see the discussion of the disanalogy of shape concepts with colour concepts in Wright (1993). In the case of shape, for example, the conditions *C* have to record the fact that the object does not change its shape during the observation. But this is to presume that facts about shapes can be determined

third condition is that there is no other competing explanation to what makes true our judgements about a given subject matter.

Wright applies this general conception of judgement-dependency to one's beliefs about one's own intentional states. On his view, such beliefs determine constitutively the truth-values of certain judgements involving intentional state concepts: "subjects' best opinions determine, rather than reflect what is true to say about their intentional states" (Wright 1989b: 250). Due to certain complications, which will be outlined shortly, he presents his biconditionals not as basic equations, but as *provisional* equations, which have the following form:

$C(\mathrm{X}) \rightarrow (\mathrm{X}$ believes that $p \leftrightarrow p)$

The basic equation cannot be applied if one cannot rule out beforehand that the obtaining of the conditions *C* does not have an effect on the truth-value of *p*. If this cannot be ruled out, as in the case of self-knowledge, then we can only consider what makes *p* true under the conditions *C* and say nothing about what happens in non-*C*-conditions. Then, if our judgements are extension-determining, they are only partially so, for it might be that they are not extension-determining under other conditions.

Wright (1989b: 252) presents the following *a priori* credible provisional biconditional for the case of intention:

"C(Jones) → (Jones believes he intends to φ ↔ Jones intends to φ)"

From this, we can see that in certain conditions, intending is determined by one's judging that one is intending. In this regard, the connection with self-directed judgements is written into the very concept of intention itself. What are the conditions *C* in the case of intention? Wright (1989b: 251) proposes that the *C*-conditions have to include that the subject possesses appropriate concepts, is attentive to their intentions and is not self-deceived. As concerns the third condition, then, Wright suggests that the condition of excluding self-deception is by its nature a "positive-presumptive" condition. If we have no evidence for self-deception, then due to the positive-presumptive nature of intentional ascriptions, one has an *a priori* justification to hold that the subject is not self-deceived.

The positive-presumptiveness of the *C*-condition thus makes the biconditional *defeasibly* extension-determining: the judgement will be overridden if there is evidence for self-deception. According to Wright (1989b: 253), such evidence could be made available by cases where "the subject's behaviour clashes with the ingredients

independently from our best opinions. In other words, since facts about an object's shape play a role in determining X's beliefs about this object, we cannot use those beliefs as determinants of facts about an object's shape (cf. Wright 1993: 80).

in his/her self-conception". By the latter, Wright presumably means one's beliefs or opinions concerning one's own intentions. Ordinarily, however, one is not deceived about one's own mind and thus one's opinions and avowals about one's intentions have an authoritative status. Wright (1989a: 632) stresses that the authority of avowals is constitutive with respect to identifying one's intentional states (cf. also Wright 1987: 402). That is, when trying to find out what the intentions and other intentional states of other people are, we have to consider their avowals.

Given this conception of judgements about one's intentions, it may seem that there is no room for asking what enables first-person authority. However, even if it is the case that under certain conditions, one's opinions about one's intentions constitute the fact that one has such intentions, it could still be asked in virtue of what do one's opinions about one's intentions acquire such an authoritative status. Wright has actually provided some hints towards an explanation, although his remarks on this are rather scarce. He points out that when we compare two interpretations of a subject's mental states, where the first one considers all evidence apart from the subject's avowal and the other one also takes the avowal into account, the latter is invariably "enormously more illuminating" (Wright 1989a: 632). The advantage of the latter interpretation is based on "the contingency that we are, each of us, ceaselessly but – on the proposed conception – subcognitively moved to opinions concerning our own intentional states which will indeed give good service to others in their attempt to understand us" (Wright 1989a: 632). The appeal to subpersonal processes does not conflict with the point that one's avowals determine truths about one's intentions, when we view this as an explanation of the enabling conditions of avowals. A conflict would ensue if it were assumed that there are certain subpersonal facts about intentions that one's opinions detect. But this assumption is not necessary: the subpersonal facts should not be construed as facts about intentions. Subpersonal processing can also be used to explain why avowals need not have accessible reasons and thus no substantial personal-level epistemology, so this approach remains consonant with the general insubstantialist line.[2]

Goldberg and Pessin (1997) construe Wright's appeal to subpersonal processing as his explanation for the presumed fact that our avowals cohere with our behaviour: this "is a matter of subcognitive processing *moving* a person to form first-person opinions that as a matter of fact cohere with her own behaviors" (Goldberg and Pessin 1997: 235). Although this overlooks some subtleties in Wright's own proposal as outlined above, it is compatible with Wright's position. Given that understanding a person is a matter of coming to a coherent conception

2. See also Wright (1989b: 253, my italics): "When possession of a certain intention is an aspect of a self-conception that *coheres* well enough both internally and with the subject's behaviour, there is nothing else that makes it true that the intention is indeed possessed."

concerning this person that incorporates her avowals as well as the rest of her behaviour, the avowals which the subject is "moved to" by the subpersonal processes must be such that they also cohere with her behaviour and thereby facilitate her interpretation by others.

To summarise Wright's explication of first-person judgements: the authority of first-person judgements does not require explanation as a substantial cognitive achievement. It is instead a matter of the conceptual features of these judgements, being based on the way in which the truth for such judgements is determined. Namely, one's best opinions about one's intentional states determine what intentional states one has, given certain conditions (including the condition that one is not self-deceived). This point about determination is underwritten by the contingent fact that one's judgements about one's own intentional states flow from one's subcognitive processes. Both of these aspects make Wright's approach relevant for present purposes. They lay the ground for further discussions in this chapter.

In general, I am sympathetic to the idea that truths about our mental states are judgement-dependent, if one has a substantive account of truth, or, that facts about mental states are dependent on our best judgements, if one prefers to talk about facts instead of truths. On this approach, intending is not an autonomous hidden mental event; on the contrary, it is constituted by one's judgement that one intends. Of course, according to the ascription theory, intending φ is not constituted by one's best beliefs under conditions *C*, but rather by the fact that intending φ is canonically ascribable. However, as we will see later, in the case of self-knowledge, canonical ascribability and one's self-ascription are closely connected. But one should not let this similarity hide a crucial difference. Namely, in Wright's account, the extension-determining facts are first-personal judgements. This seems to have the outcome that first-person and third-person concepts of intention differ (see also Holton 1993: 302–303). There is one story to be told about me having an intention and another story told about my intentions from the third-person perspective. Presumably, the third-person concept of intention is extension-tracking and not extension-determining for Wright. It tracks the extension that is fixed by first-person sincere self-reflective judgements. Nothing like this is implicated in the ascription theory, which operates under the constraint that first- and third-person concepts are the same.

9.3 First-person operationalism

Besides the already discussed constitutivist idea, there is another compelling aspect to Wright's account. It is the thought that first-person authority can be explained by the positive-presumptiveness of our avowals. We are by default

justified in assuming that the subject is not undergoing self-deception. The idea of positive-presumptiveness can be generalised to make it central to the approach that I call 'first-person operationalism', to use a term from Dennett (1991a: 132). Namely, first-person operationalism, or FPO for short, is a certain approach to the first person which is characterised by the following dictum:

> (FPO) If one says (avows, judges, reports, etc.) that one has a mental state (with a certain content), then one has this mental state (with this content)

For Dennett, first-person operationalism is the view that denies "the possibility in principle of consciousness of a stimulus in the absence of subject's belief in that consciousness" (Dennett 1991a: 132; see also Rosenthal 1994). Dennett rejects the appearance-reality distinction in the case of self-consciousness and is opposed to unverifiable and unmanifestable conscious happenings, so FPO is a path to take, since there are no deeper mental facts that can override a subject's judgement. Wright's approach must be driven by basically the same idea. Intending φ is constituted by a subject's judgements under conditions *C*; there are no deeper hidden facts.

In my view, the thesis of first person operationalism is one of the working principles of folk psychology. Indeed, it explains first-person authority in a very simple manner. First-person authority is a relational concept. Someone acts as an authority if others regard them in this way. A person is conferred presumptive authority with respect to their own mind not because of the special nature of mental properties or events, but because of the fact that normally one's reports are in accord with one's actions and other relevant facts. We do not need to subscribe to any deep metaphysics about minds in order to follow the dictum of FPO. In addition, following this principle does not commit one to any inflationary higher-order mental properties and states. One can follow the folk-psychological form of life and say that one has beliefs and desires, but it is still open to one to give an interpretivist account of what it is to have beliefs and desires. In this respect, being underwritten by a principle of our common-sense psychology, first-person authority has a social basis. It is neither mysterious nor metaphysical.

FPO seems to exclude the possibility of rational disagreement with others about their minds. However, we should not forget that it holds only defeasibly, given the positive presumption that the subject is not self-deceived. On the presumption that there is no countervailing evidence, FPO is the principle to follow. In terms of the ascription theory, this amounts to the claim that in the interpersonal context, the first-person ascription is positively presumed to be canonical, if one is to choose between the first- and the third-person ascriptions. If the choice is between conflicting first-person actual ascriptions, then the situation becomes

more complicated when we approach it from the viewpoint of the actual choice situation. From the canonical point of view, the matter is straightforward: if there is a genuine conflict, then at least one of the ascriptions is not canonical. In the situation where one has to choose between two conflicting self-ascriptions, one has to look for the cause of this conflict, and this may indicate that the positive-presumptiveness condition is not fulfilled.

There remains the task of explaining why FPO normally works. The answer cannot be just that avowals constitute mental states and that it is a grammatical fact that this is so, since then the origin of this fact would remain mysterious. We also need to explain why avowals have such a special status and, as we saw in the previous section, Wright has given us a clue in his brief remark about the subcognitive mechanism. Needless to say, this requires greater specification and a place in a larger framework. Otherwise, waving a hand towards some mechanisms seems *ad hoc*, especially given that the success of FPO can also be given a shallow explanation in terms of FPO being a constitutive presumption of interpretation (see also Goldberg and Pessin 1997: 236–237).[3]

What would such a shallow account look like? It would be a broadly Davidsonian approach in which authority is granted to the first person by interpreters who proceed from considerations of charity, and not the result of some internal configuration. The point is that the assumption of first-person authority is a precondition of interpretation, for in Davidson's view, as we saw in Chapter 3, interpretation proceeds from the sentences the subject holds true. Since what is meant by such utterances depends on both the subject's beliefs and what the subject means, we can deduce the latter when we assume that the subject has true beliefs about his surroundings. This procedure would be undercut if the subject was constantly mistaken about what she holds true and we could not interpret her utterances.

> It makes no sense … to wonder whether the speaker is generally getting things wrong. His behavior may simply not be interpretable. But if it is, then what his words mean is (generally) what he intends them to mean. … There is a presumption – an unavoidable presumption built into the nature of interpretation – that the speaker usually knows what he means. So there is a presumption that if he knows that he holds a sentence true, he knows what he believes.
>
> (Davidson 1984b: 111)

This gives us one explanation of the origins of the assumption of first-person authority. A similar account can also be given of the coherence between one's

---

3. By stressing only the constitutive role of avowals, Wright actually comes close to this shallow view in his earlier paper (see Wright 1987: 401–402).

behaviour and judgements about mental states. Such coherence can be viewed as another precondition of interpretation, projected by interpreters. Goldberg and Pessin have expressed this idea vividly:

> [T]hat a subject's first-person opinions cohere with her own behavior is not in fact a substantial achievement after all. On the contrary, these opinions *have to* cohere with her behavior (more of less) on pain of not being recognized as first-person opinions at all. And to say that one's first-person opinions *have to* cohere with one's behavior here means that we (the interpreters of that person) *impose* this (by and large) coherence on the person's behavior when we interpret them. We *impose* it; we do not *discover* it. (Goldberg and Pessin 1997: 237)

It is not altogether obvious whether Davidson himself shares the latter view as well. It is true that he stresses the role of coherence in interpretation, and suggests that to ascribe incoherent thoughts to someone amounts to declaring that they are not thinkers at all. However, Davidson does not hold onto pure ascriptivism and presumably would not agree with the implication that coherence is only in the eye of the beholder. For this reason, we should call an account that explains both first-person authority and the coherence between one's first-person avowals and behaviour, a 'Davidsonian approach'.

One could try to conjoin this broadly Davidsonian approach with the ascription theory, but I think that the result would be crucially incomplete. In the resulting account, FPO would be a folk principle that guides interpretation and, as a precondition of ascription, it is assumed by default that it holds. This assumption can be reinforced with the coherence of one's behaviour with one's self-ascriptions. However, on the proposed account, the reinforcement is nothing substantial, when it is taken as a result of the fact that the interpretation of the behaviour involves maxmising its coherence.

At this point we should distinguish two kinds of relations for which coherence might be at issue: the coherence between one's mental states and behaviour and the coherence between one's self-ascriptions and behaviour. Although my account, given at the end of this chapter, actually provides a unifying explanation for these two kinds of relations, it is important to distinguish them for purposes of clarity. In the former case, mental states that can be ascribed are chosen partially in light of behavioural considerations and hence it can be expected that, as a result, they come to cohere with behaviour. However, even then, we should explain why we let our interpretations be guided by the assumption of coherence. The other case is already *prima facie* different: there cannot even be a question of whether the coherence between a subject's actual acts of self-ascription and behaviour is imposed by interpreters. It is rather an observable and a contingent matter of fact that one's avowals and judgements about one's own mental states are

normally consistent with how one acts. It is the latter kind of coherence relation that has to be explained when one explains the FPO principle. What we need to do is to explain the contingent basis of that constitutive principle. One way to do this is to present an account of the subpersonal machinery underpinning acts of self-ascription. This is the topic of the following sections.

9.4 The idea of a mechanism

Accordingly, let us assume that one has to invoke a subpersonal mechanism in order to explain the general reliability of self-knowledge and the coherence of self-ascriptions with behaviour. The idea of postulating a mechanism that produces self-ascriptions is a reliabilist idea. I begin by outlining one way in which such a mechanism can be conceived and then I discuss whether some such approach is open to the interpretivist.

It is not uncommon to invoke the workings of some reliable mechanism to explain self-knowledge. This picture takes self-knowledge to be a second-order belief about one's first-order states, whereas the second-order belief is taken to be the output of a reliable process. The reliabilist approach states that the second-order states are justified in virtue of the general reliability of the mechanism that produces them. But the mechanism itself is not normally available to a subject as the subject's reason for holding the second-order belief. For the subject, this second-order belief seems to arise out of nowhere. In this sense, first-person epistemology is cognitively insubstantial. The reliabilist view is nicely captured by Sydney Shoemaker:

> Our minds are so constituted, or our brains are so wired, that, for a wide range of mental states, one's being in a certain mental state produces in one, under certain conditions, the belief that one is in that mental state. … The beliefs thus produced will count as knowledge, not because of the quantity or quality of evidence on which they are based (for they are based on no evidence), but because of the reliability of the mechanism by which they are produced. (Shoemaker 1994: 268)

Can this picture be reconciled with interpretivism? We need to discuss this more closely. There could also be a discrepancy in Wright's case. If we postulate a mechanism to process mental states, there has to be a judgement-independent internal state to serve as an input for the mechanism which computes and outputs a second-order belief. However, as we saw, on Wright's account, first-order states (e.g., intentions) are judgement-dependent.

There is a general problem of the compatibility of personal and subpersonal explanations. Applied to the present case, the problem is that if we take the

mechanism hypothesis as an explanation for the origin of second-order beliefs, then this seems to make all personal-level constitutive accounts of self-knowledge either simply redundant or derivative from the fact that some such mechanism is present. This is a large topic, and I can only be brief here. As already noted, one option is to argue that personal-level constitutive accounts like Wright's view and the ascription theory are not in competition with subpersonal accounts, since they do not apply to the same thing. One is a personal-level story about mind and the other is a subpersonal account of the enabling conditions of the personal-level story.

What the constitutive account, and here I have especially the ascription theory in mind, cannot allow is a model that postulates mental states that are independent from the ascription. A mechanism that produces the second-order state as an effect of having the first-order state as an input, presumes such mental states and processes that have no constitutive connection with interpretation. Hence, we cannot just adopt the outlined reliabilist picture without alteration. A modified reliabilist view that we can accept is presented in the next section. It is a model of self-knowledge that has a slot for subpersonal mechanisms, but which does not make interpretation redundant.

One might attempt to uphold the compatibility of the idea of a mechanism and interpretivism by claiming that the possession of both mental states and the mechanism that computes them depend on interpretation. However, then it seems difficult to square the fact that mental states are personal-level states with the fact that the mechanism is ascribed at the subpersonal level. For how could personal-level states that are ascribable to the whole person and not to its parts enter into the subpersonal mechanism? But if we assume that the first- and second-order states are subpersonal states, then self-knowledge comes down to knowing first-order subpersonal states in virtue of being in second-order subpersonal states. This clashes with the idea that the states that are accessible to us in the case of self-knowledge are the same kind of personal-level states that we employ in everyday life when we ascribe them to others.

9.5 The ascriptionist model of self-knowledge

In this section, I will construct a model of self-knowledge that respects the principle of first-person operationalism, and provides this normative constraint with some basis in the subpersonal mechanism. This model does not presume the existence of mental states and contents that are independent of the ascription.

Imagine a subpersonal process computing sensory input from either the outer or inner environment. From time to time, one's subpersonal processes trigger self-ascriptions. Sometimes additional external prompts are necessary. In that case, a

self-ascription is triggered that is usually in accord with all the relevant ascription sources for the particular content that is ascribed. Take, for example, a subject's actual self-ascription of the belief that *p*. This self-ascription itself does not yet settle whether the subject has a belief that *p* or not. In the ascription theory, this is constituted by the canonical ascription. If the belief that *p* is canonically ascribable to the subject, then she believes that *p*. This model also allows for the case of mistaken self-ascription. This is the case when one ascribes to oneself a belief that canonically is not properly ascribable to one. However, I think that such cases, although possible, are not very common. The reason for this is that we are blessed with reliable subpersonal mechanisms that make it so that whenever a state with a conceptual content, e.g., a belief that *p* is canonically ascribable, if a subject is suitably prompted, barring other impediments, the self-ascription of the belief that *p* is triggered.[4] This is the deeper basis of first-person operationalism. Hence, since a subject's own ascriptions are usually proven canonical, it is safe to assume that they are so. Of course, such an assumption can only be presumptive, but so too is the FPO principle. In the present model, the mechanism is not postulated between the first-order and the second-order states; instead it is one that enables agreement between the self-ascription and what is canonically ascribable. This is the crucial difference from the reliabilist model. By conceiving the mechanism in a different way, we can avoid postulating ascription-independent mental processes.

It is important to be clear what kind of mechanism is involved in this model. First, the mechanism need not and presumably is not a cognitive unit. Rather, by talking about 'the mechanism', I intend to refer to a manifold of structural units with varying complexity. Second, we have to clarify which states are involved in this mechanism. It was already pointed out that these are not the first-order and the second-order mental states themselves. Instead, the mechanism ensures that there is an agreement between what is canonically ascribable and what is self-ascribed. This agreement can be viewed as the indirect result of the mechanism, for the mechanism has a direct effect only on the triggering of a suitable self-ascription in a given setting. The ascription will be suitable for the given setting since one's subpersonal processes underpin and maintain the harmony, broadly conceived, between the environment and our reactions to it. This is not to say that each environmental event entails a specific reaction; it is rather the point that each reaction can be explained in terms of facts included in the ascription sources, including the enviromental stimuli. Hence, we can say that these various subpersonal processes that maintain one's behaviour in one's environmental setting undergird the common-sense idea that the subject's mental states are related to

4. This does not apply for states with non-conceptual content, for one cannot self-ascribe non-conceptual content to oneself. See the discussion in Chapter 8.

what he sees and to what he does in a harmonious and coherent manner. Making proper self-ascriptions is a part of this overarching harmonious ensemble. That is, reliable subpersonal mechanisms underpin the coherence between what one says about one's own mind and how one acts in a given setting.

As explained in Chapter 5, one condition for a mental state to be canonically ascribable is that ascribing such a mental state is maximally coherent with the ascription sources (or approximates such coherence). We can see that both the self-ascription and the canonical ascription move towards coherence with the ascription sources. In the former case, this is due to the mechanism that triggers proper self-ascriptions; in the latter case, it is due to the interpretation that conforms to the conditions of canonicality. Both ascriptions converge in being coherent with the sources and this allows us to privilege self-ascriptions as presumptive candidates for canonical ascriptions.

The general model is depicted in Figure 2.

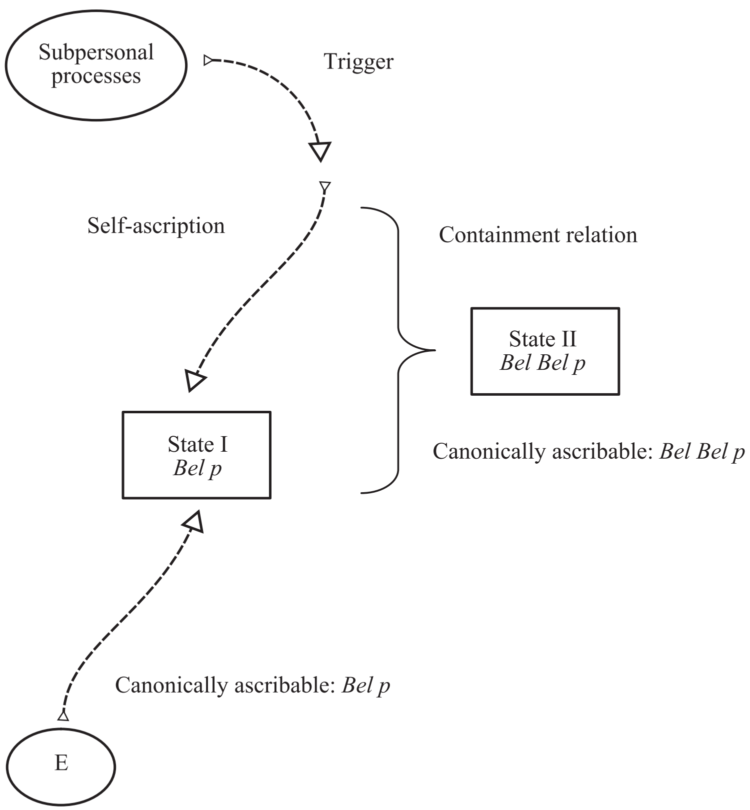


Figure 2. Ascriptionist model of self-knowledge.

In Figure 2, we see the triggering of an act of self-ascription by subpersonal processes. It can be seen from the line below that the same state is canonically ascribable as well. The canonical ascription is based on evidence pictured by an encircled 'E', whereas the self-ascription is not normally based on the ascription sources. On the right side of the figure, there is a second-order state, which is a belief about one's belief. As in the case of the first-order state, the possession of that state is constituted by the canonical ascription and not by the self-ascription. However, the fact that one has made the actual self-ascription of the belief that *p* to oneself, normally makes it canonically ascribable that one has a second order belief, the belief that one believes that *p*.[5] This does not imply that the person herself would have to make this second-order self-ascription. Some people (e.g., Burge 1988; Heil 1988 and Shoemaker 1994) have exploited the containment of the first-order content in the second-order content to argue for first-person authority or even infallibility. However, containment has no special role in backing up authority in the present model; but it does have a role in the justification of self-knowledge, and I will return to this issue shortly.

In the model, there are subpersonal processes leading to acts of self-ascription. Using the notion of protocontent that was introduced in the previous chapter, we can say that by way of having brain processes, the subject also has protocontent. In the present model of self-ascription, subpersonal brain processes are causally efficacious in producing the self-ascription and one might want to say that protocontent plays a causal role. At the present level of detail, however, it is difficult if not impossible to distinguish between those subpersonal processes that play a part in triggering self-ascriptions and appear to a subject phenomenally and those that play a part in triggering but do not appear. Therefore, I have not made the place of protocontent explicit in the model in Figure 2. An alternative would be to replace the label 'subpersonal processes' with that of 'protocontent', but this would entail that protocontent triggers acts of self-ascription. This would in turn presume that we can distinguish between different stages or tokens of protocontent, so that some protocontents trigger ascriptions of a certain kind, some trigger ascriptions of a different kind and some protocontents do not trigger ascriptions at all. However, as noted in the previous chapter, I cannot make sense of such temporal discrimination in protocontent. In addition, this would raise difficult questions concerning the role of other subpersonal processes in the production of self-ascription. Is it the case that only those subpersonal processes

5. In the case of other kinds of mental states like desires, the second-order thought about desire is not itself a second-order desire. In that case, I would say that the fact that one has made the actual self-ascription of a desire to oneself makes it canonically ascribable that one has a second order thought about that desire.

trigger ascriptions that also appear to the subject? My contention is that it is more reasonable to say that the act of ascription is triggered by subpersonal processes and some such subpersonal processes underpin what I call 'protocontent', but we cannot make the more specific claim that only those subpersonal processes that manifest phenomenally lead to self-ascription.

It is also important to stress that protocontent is neither evidentially nor justificationally related to the self-ascribed state. This was already discussed in the previous chapter, and I provide a short recapitulation, for it is relevant to the epistemic status of self-knowledge. Protocontent is not evidentially related to mental states due to its peculiar nature. The first reason is that there is no way to match it up with (conceptual) contents that it would justify. We cannot point to a slice of protocontent and say that this is the non-conceptual ground for conceptual self-ascription. The second reason is that this would allow the existence of a very peculiar sort of private evidence that only a person herself would have for her own self-ascription. If that were so, there would be a crucial difference in the ascription from the first-person and the third-person point of view. The set of evidence available to the first person would then also include protocontent. This is exactly what interpretivism denies: on the interpretivist view, all evidence is in principle public.[6] To deny that protocontent is private evidence, however, is not to deny the difference between the first person and the third person. On the face of it, interpretivism in general does not seem to line up happily with the idea of an asymmetry between first-person and third-person access to the mind. However, in my account, we can say that there is something extra in the first person case, namely protocontent, which adds the special *feel* to the apprehension of the world from the first-person perspective, but – given the interpretivist standpoint – it does not play a role in the constitution of content-possession.

This is the insubstantial element in my account. There is something like to have protocontent, but this cannot be said, for it is ineffable; and its relation to content is not evidential. Self-ascriptions would seem to arise out of nothing in the sense that they are not usually the product of a conscious weighing of evidence and pondering. That is why self-knowledge seems direct. This does not entail that the results of such seemingly baseless ascriptions are not in harmony with the

---

6. These are some of the reasons why I cannot accept an account like Goldman's model of self-ascription in which the qualia representation is matched with the corresponding category representation and according to which this matching is effected on the basis of qualia-detection alone (Goldman 1993: 20). Even if we concede to Goldman that there is such a qualia-detection mechanism, the mere qualia-detection would not be sufficient for the matching, since there is much more besides one's qualia that matters in singling out the propery for categorisation, for instance, some background information.

relevant evidence. As stated, such harmony is guaranteed by the reliable mechanism. Neither does it follow that there cannot be a justification for the ascription. The justification is forthcoming whenever it can be constructed. The idea is that a similar justification can in principle be composed for the self-ascription and the canonical ascription. If one's theory of knowledge takes the personal or internal justification to be a condition of knowledge, then the canonical self-ascription is also self-knowledge, for there is a justification available for it in terms of the relevant ascription sources and their common-sense rules. But it is not necessary that the subject himself is able to compose and state that justification. If one's theory of knowledge is externalist, then one can explain that self-ascription qualifies as knowledge since it is subserved by reliable mechanisms.

I would like to anticipate an objection. It might be objected that if one were asked why one thinks that one believes such and such, the answer would normally be, "Because I just have this belief". One would not normally say that considering such and such circumstances, it makes most sense to say about me that I believe such and such.

An interpetivist can give two compatible answers to this objection. First: the folk-psychological story about what one would say is the object-level story. When one talks about the possibility of constructing a *post factum* justification for ascription, one is talking at the metalevel where one can draw a distinction between what justification is in folk psychology (what people usually say) and what the philosophically acceptable justification is. The second answer exploits the containment of first-order states in second-order states. When one says that one believes that *p*, one ascribes oneself the belief that *p*, which in turn is a higher-order state of believing that one believes that *p*. If we look for the justification of the possession of the belief that one believes that *p*, we may, among other facts, cite the subject's self-ascription of the belief that *p*. But when we look at what is involved in believing that *p*, the interpretivist story is that this must be the most apt thing for the subject to do given his situation. The ascription-independent mental state is nowhere involved in that answer.

Finally, I should note that what has been presented here is a model for basic self-knowledge. By this, I mean a plain judgement about one's own mental states that does not involve complex reflection on one's motives and pondering of various kinds of evidence.[7] This model can be extended to cover more complex kinds of self-knowledge. The most immediate addition would be to say that

7. For a role of deliberation in self-knowledge, see Moran (2001). Tanney (2002) discusses the extent of construction in self-cognition. For a recent cognitivist conception of self-knowledge that distinguishes between the primitive scanning mechanism and the theory of mind module that is involved in reasoning about one's own mind, see Nichols and Stich (2003).

what is sometimes triggered is not an act of self-ascription, but the consideration and weighing of evidence that precedes self-ascription. However, in some cases, self-ascription can involve making new decisions about how one should think or forming a complex representation of oneself. Sometimes, the act of self-ascription or becoming conscious about how one feels may influence the interpretation of oneself. An initial ascription may thus be modified by further reflections. There are no principled reasons why such relatively complex forms of self-cognition cannot be accommodated in the present model, but this would require more extensive elaboration than I can provide in the present book.[8]

The account presented in this chapter proceeds from the observation that the authority of self-knowledge is based on a folk-psychological presumption. This presumption derives from our practical observation that people are usually correct about their mental states. Following a clue provided by Wright, I presented a model in which a reliable subpersonal mechanism ensures that we make proper self-ascriptions, which can presumptively be regarded as candidates for canonicality. In this way, the first-person privilege receives a twofold backing. First, as expressed by the FPO principle, it is part of our folk psychology that presumptively the first person's opinions have an authoritate status. Second, the model provided displayed that this principle is not *sui generis* or a shallow fact of the "grammar" of intentional ascription, but has a basis in our cognitive machinery.

The inability to devise a satisfactory account of self-knowledge seemed to be one of the most serious obstacles to interpretivism. Therefore it was highly desirable to come up with an ascriptionist account of self-knowledge that respects our most basic intuitions. Although the basic model should be extended in various ways and much remains to be said about self-knowledge from the interpretivist perspective, I think that we now have reason to believe that the ascription theorist can carry this burden.

8. Another class of states whose self-ascription would require modification of the model is the class of partial beliefs (having a degree of confidence that *p*). The question of the relation between partial beliefs and flat-out beliefs is a large topic of its own. For some interesting developments, see Dennett (1978: Chapter 16) and Frankish (2004).

CHAPTER 10

Conclusion

My main aspiration for this book was to put interpretivism back on the map, that is, to show that it is indeed a position that deserves to be taken seriously. As a part of this venture, I argued that several alternative positions on what it is to have a mind are prone to objections on different counts. This formed the bulk of Chapter 2. In that chapter, I proceeded from the distinction between entities that are only deflationary or pleonastic and natural existents that are entities in the inflationary sense. One reason for using such a distinction was to explicate the commitments of various positions on the mind-body relation. Another reason was to draw attention to the point that a justification for assuming that there are mental entities in the inflationary sense is lacking, for all we are committed to in using mental vocabulary when making common-sense ascriptions is just pleonastic mental entities. The latter, however, do not pose insuperable metaphysical worries. The task is rather to explain the nature and the possession conditions of pleonastic mental entities. And this can be accomplished by explaining the meaning of mental terms and explicating the conditions on which mental terms are applicable. The central idea is thus that the possession of mental properties comes down to a metaphysically far less problematic matter of being describable by mental terms, more specifically, being subject to interpretation in mental vocabulary.

Metaphysical matters came to the fore again in Chapter 7, where I attempted to make sense of mental causation from an interpretivist perspective. There, I again made use of the distinction between deflationary and inflationary entities and argued that in one sense mental entities play no causal role and in another sense they do. Namely, mental entities as pleonastic entities can have no causal efficacy, which belongs to the elements of the natural basis. Thus, they do not take part in causal relations. However, mental entities can be causally relevant, in the sense explained in the chapter, and play a major explanatory role, especially in folk psychology. This explanatory role can be retained even if mental entities are not efficacious. The distinction between causal efficacy and relevance enables us to accommodate the causal commitments of folk psychology without incurring severe metaphysical costs.

My version of interpretivism was inspired by the work of Dennett and Davidson. I have discussed their views and related my own claims to these views throughout the book. However, the main place where their approaches are scrutinised is

Chapter 3. In that chapter, I distinguished between various possible versions of interpretivism and gave an overview of Davidson's and Dennett's approaches, which are quite different on close inspection. My overview was also partly critical and I pointed out several areas where my approach differs from theirs.

Next, I started to pave the way for my own version of interpretivism: the ascription theory. Chapter 4 settled some preliminary matters concerning the status and the meaning of mental terms as well as the role of folk psychology for the ascription theory. Folk-psychological understanding of the mental is important for the ascription theory, since it provides those principles and interconnections on the basis of which mental terms are applied. Along the way, I also distinguished my approach from common-sense functionalism about mental terms and rebutted attempts to construe disagreements over the nature of the mental as disagreements over the correct theory of reference for mental terms.

Chapter 5 presented the ascription theory, construed as an account of what it is to *have* mental states with contents, instead of an account of the *nature* of mental states. This theory explicates the possession of mental properties in terms of their canonical ascribability. Canonicality sets a certain standard for an interpretation in order to ensure that not every arbitrary ascription would count as determining what mental properties one has. On the other hand, this should in principle be accessible for ordinary interpreters. I also explained the requirements of canonicality (the coherence condition as well as the revisability constraint) and elaborated my stance on the indeterminacy of the mental, which is an issue that has received diverse treatment among interpretivists.

In Chapter 6, I went on the defensive mode by countering objections that might be addressed to the ascription theory. These objections concern the apparently problematic circularity of the account, the problem of regress, liberalism and observer-dependence. The worry that the account does not explain intrinsic intentionality was answered as well. I finished the chapter by bringing out considerations that might lead one to support the interpretivist position and worked them into a sketch of an argument for interpretivism.

Chapters 8 and 9 discussed the applications of the ascription theory to perception and self-knowledge, respectively. Due to their specificity, both areas present different challenges to interpretivism. One question is whether an interpretivist position is able to take the phenomenal aspects of perception into account and another widespread worry is that interpretivism fails to do justice to first-person authority. In both cases, I sought to defend interpretivism. The notion of a nonconceptual content of perception and the mechanisms underlying self-knowledge received some attention as well.

I do not believe that the primary task for a *philosophical* approach to the mind is to explain how mind functions or how the brain works. This is a task for

the cognitive sciences. In line with this outlook, I made use of one philosophical construction, protocontent, whose structure and functions are something that might be explained by scientific study, but which is a black box from the folk-psychological point of view. Instead of making fine-grained empirical hypotheses on its nature, I rather employ this construction to make sense of some philosophical issues concerning the phenomenality of perception and self-knowledge.

Several topics, especially those in later chapters, deserve more extended treatment, but in order to keep this book to a manageable length, I have mostly concentrated on how the issues would appear from the perspective of the ascription theory rather than elaborating the themes in their own right. There are also some topics that were discussed only briefly and in passing, but which are relevant for giving a more complete background to the ascription theory. Two such tasks spring to mind immediately: the development of an account of meaning that can support the ascription theory and giving a more detailed account of the relation between subpersonal explanation and folk psychology. I hope that I will have an opportunity to expand my discussion of these matters on some future occasion.

Looking back on the book, a certain pattern seems to emerge. My overall approach has been guided by the wish to provide an alternative perspective on mainstream problems in the philosophy of mind. As I think that several philosophical worries are of our own making, I have tried to deflate several such worries, as in the case of the problem of mental causation and the ontology of mind, by examining them from a perspective from which traditional concerns do not seem insurmountable anymore. In view of the fact that there is such an alternative picture, where these problems no longer arise, one can make a case for them not being real problems in the first place.

References

Anscombe, G. E. M. 1981. "Causality and determination." In *The Collected Philosophical Papers of G. E. M. Anscombe, II: Metaphysics and the Philosophy of Mind*, G. E. M. Anscombe, 133–147. Oxford: Basil Blackwell.

Antony, L. 2003. "Who's afraid of disjunctive properties?" *Philosophical Issues* 12: 1–21.

Armstrong, D. M. 1968. *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Armstrong, D. M. 1978. *Nominalism and Realism*. Cambridge: Cambridge University Press.

Armstrong, D. M. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press.

Bach, K. 1982. "De Re belief and methodological solipsism." In *Thought and Object: Essays on Intentionality*, A. Woodfield (ed.), 121–151. Oxford: Oxford University Press.

Baker, L. R. 1989. "Instrumental intentionality." *Philosophy of Science* 56: 303–316.

Baker, L. R. 1995. *Explaining Attitudes: A Practical Approach to the Mind*. Cambridge: Cambridge University Press.

Baker, L. R. 2001. "Practical realism defended: replies to critics." In *Explaining Beliefs: Lynne Rudder Baker and Her Critics*, A. Meijers (ed.), 183–217. Stanford, CA: CSLI Publications.

Baker, L. R. 2003. "Belief ascription and the illusion of depth." *Facta Philosophica* 5: 183–201.

Baldwin, T. 1993. "Two types of naturalism." *Proceedings of the British Academy* 80: 171–199.

Barber, A. 1998. "The pleonasticity of talk about concepts." *Philosophical Studies* 89: 53–86.

Bermúdez, J. L. 1994. "Peacocke's argument against the autonomy of nonconceptual representational content." *Mind and Language* 9: 402–418.

Bermúdez, J. L. 1995. "Nonconceptual content: from perceptual experience to subpersonal computational states." *Mind and Language* 10: 333–369.

Bickle, J. 2002. "Multiple realizability." *The Stanford Encyclopedia of Philosophy* (Fall 2002 Edition), E. N. Zalta (ed.).

Bickle, J. 2003. *Philosophy and Neuroscience. A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Binderup, L. 2004. "Is there a single true theory of properties?" *Res Cogitans* 1: 48–90.

Bishop, M. and Stich, S. 1998. "The flight to reference, or how not to make progress in the philosophy of science." *Philosophy of Science* 65: 33–49.

Blackburn, S. 1994. *The Oxford Dictionary of Philosophy*. Oxford: Oxford University Press.

Block, N. 1978. "Troubles with functionalism." *Minnesota Studies in the Philosophy of Science* 9: 261–325.

Block, N. 1980. "What is functionalism?" In *Readings in the Philosophy of Psychology*, vol. 1, N. Block (ed.), 171–184. Cambridge, MA: Harvard University Press.

Block, N. 1981. "Psychologism and behaviorism." *Philosophical Review* 90: 5–43.

Block, N. 1995. "On a confusion about a function of consciousness." *Behavioral and Brain* Sciences 18: 227–247.

Block, N. 1997. "Anti-reductionism slaps back." *Philosophical Perspectives* 11: 107–132.

Block, N. 2003a. "Do causal powers drain away?" *Philosophy and Phenomenological Research* 67: 133–150.
Block, N. 2003b. "Mental paint." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, M. Hahn and B. Ramberg (eds), 165–200. Cambridge, MA: MIT Press.
Block, N. and Stalnaker, R. 1999. "Conceptual analysis, dualism, and the explanatory gap." *Philosophical Review* 108: 1–46.
Boghossian, P. 1989. "Content and self-knowledge." *Philosophical Topics* 17: 5–26.
Bonevac, D. 2002. "Sellars vs. the given." *Philosophy and Phenomenological Research* 64: 1–30.
Bortolotti, L. 2004. "Can we interpret irrational behavior?" *Behavior and Philosophy* 32: 359–375.
Bortolotti, L. 2005. "Delusions and the background of rationality." *Mind and Language* 20: 189–208.
Braddon-Mitchell, D. and Jackson, F. 1996. *Philosophy of Mind and Cognition*. Oxford: Blackwell.
Brandom, R. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.
Brewer, B. 1999. *Perception and Reason*. Oxford: Oxford University Press.
Burge, T. 1979. "Individualism and the mental." *Midwest Studies in Philosophy* 4: 73–121.
Burge, T. 1988. "Individualism and self-knowledge." *Journal of Philosophy* 85: 649–663.
Burge, T. 1993. "Mind-body causation and explanatory practice." In *Mental Causation*, J. Heil and A. Mele (eds), 97–120. Oxford: Clarendon Press.
Burwood, S., Gilbert, P. and Lennon, K. 1998. *Philosophy of Mind*. London: UCL Press.
Byrne, A. 1998. "Interpretivism." *European Review of Philosophy* 3: 199–223.
Campbell, K. 1990. *Abstract Particulars*. Oxford: Blackwell.
Carrier, M. and Mittelstrass, J. 1991. *Mind, Brain, Behavior. The Mind-Body Problem and the Philosophy of Psychology*. Berlin/New York: Walter de Gruyter.
Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
Chater, N. and Pickering, M. J. 2003. "Two realms of mental life: The non-overlap of belief ascription and the scientific study of mind and behavior." *Facta Philosophica* 5: 335–353.
Cherniak, C. 1986. *Minimal Rationality*. Cambridge, MA: The MIT Press.
Child, W. 1994. *Causality, Interpretation and the Mind*. Oxford: Clarendon Press.
Child, W. 2002. "Reply to Alvin I. Goldman." In *Simulation and Knowledge of Action*. J. Dokic and J. Proust (eds), 21–31. Amsterdam/Philadelphia: John Benjamins.
Churchland, P. 1981. "Eliminative materialism and propositional attitudes." *Journal of Philosophy* 78: 67–90.
Clark, A. 1993. *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge, MA: MIT Press.
Crane, T. 1992. "The nonconceptual content of experience." In *The Contents of Experience. Essays on Perception*, T. Crane (ed.), 136–157. Cambridge: Cambridge University Press.
Crane, T. 1998. "How to define your (mental) terms." *Inquiry* 41: 341–354.
Crane, T. 2001. *Elements of Mind*. Oxford: Oxford University Press.
Crane, T. and Mellor, D. H. 1990. "There is no question of physicalism." *Mind* 99: 185–206.
Cummins, R. 1989. *Meaning and Mental Representation*. Cambridge, MA: MIT Press.
Cussins, A. 1990. "The connectionist construction of concepts." In *The Philosophy of Artificial Intelligence*, M. Boden (ed.), 368–440. Oxford: Oxford University Press.
Davidson, D. 1963. "Actions, reasons and causes." *Journal of Philosophy* 60: 685–700.
Davidson, D. 1967a. "Causal relations." *Journal of Philosophy* 64: 691–703.
Davidson, D. 1967b. "Truth and meaning." *Synthese* 17: 304–323.

Davidson, D. 1968–69. “On saying that.” *Synthese* 19: 130–146.
Davidson, D. 1969. “How is weakness of the will possible?” In *Moral Concepts*, J. Feinberg (ed.), 93–113. Oxford: Oxford University Press.
Davidson, D. 1970. “Mental events.” In *Experience and Theory*, J. Foster and J. W. Swanson (eds), 79–101. Amherst, MA: University of Massachusetts Press.
Davidson, D. 1973a. “Radical interpretation.” *Dialectica* 27: 313–328.
Davidson, D. 1973b. “The material mind.” In *Logic, Methodology and Philosophy of Science IV*, P. Suppes et al. (eds), 709–722. Amsterdam: North Holland Publishing Company.
Davidson, D. 1973–74.“On the very idea of a conceptual scheme.” *Proceedings and Addresses of the American Philosophical Association* 47: 5–20.
Davidson, D. 1974. “Belief and the basis of meaning.” *Synthese* 27: 309–323.
Davidson, D. 1975. “Thought and talk.” In *Mind and Language*, S. Guttenplan (ed.), 7–23. Oxford: Clarendon Press.
Davidson, D. 1979. “The inscrutability of reference.” *The Southwestern Journal of Philosophy* 10: 7–19.
Davidson, D. 1980a. *Essays on Actions and Events*. Oxford: Clarendon Press.
Davidson, D. 1980b. “Toward a unified theory of meaning and action.” *Grazer Philosophische Studien* 11: 1–12.
Davidson, D. 1982. “Rational animals.” *Dialectica* 36: 317–327.
Davidson, D. 1983. “A coherence theory of truth and knowledge.” In *Kant oder Hegel*, D. Henrich (ed.), 423–438. Stuttgart: Klett-Cotta.
Davidson, D. 1984a. *Expressing Evaluations*. Lindley Lecture. Lawrence, KS: University of Kansas.
Davidson, D. 1984b. “First person authority”. *Dialectica*: 38: 101–111.
Davidson, D. 1985a. “Incoherence and irrationality.” *Dialectica* 39: 345–354.
Davidson, D. 1985b. “Replies to essays X–XII.” In *Essays on Davidson: Actions and Events*, B. Vermazen and M. Hintikka (eds), 242–252. Oxford: Oxford University Press.
Davidson, D. 1986. “A nice derangement of epitaphs.” In *Philosophical Grounds of Rationality*, R. Grandy and R. Warner (eds), 156–174. Oxford: Clarendon Press.
Davidson, D. 1987a. “Knowing one’s own mind.” *Proceedings and Addresses of the American Philosophical Association* 60: 441–458.
Davidson, D. 1987b. “Problems in the explanation of action.” In *Metaphysics and Morality*, P. Pettit, R. Sylvan and J. Norman (eds), 35–49. Oxford: Blackwell.
Davidson, D. 1989. “The myth of the subjective.” In *Relativism: Intepretation and Confrontation*, M. Krausz (ed.), 159–172. Notre Dame, IN: University of Notre Dame Press.
Davidson, D. 1990a. “Representation and interpretation.” In *Modelling the Mind*, K. A. Mohyeldin Said, W. H. Newton-Smith, R. Viale and K. V. Wilkes (eds), 13–26. Oxford: Clarendon Press.
Davidson, D. 1990b. “The structure and content of truth.” *Journal of Philosophy* 87: 279–328.
Davidson, D. 1991. “Three varieties of knowledge.” In *A. J. Ayer: Memorial Essays*, A. Phillips Griffiths (ed.), 153–166. Cambridge: Cambridge University Press.
Davidson, D. 1993. “Thinking causes.” In *Mental Causation*, J. Heil and A. Mele (eds), 3–17. Oxford: Clarendon Press.
Davidson, D. 1994a. “Radical interpretation interpreted.” *Philosophical Perspectives* 8: 121–128.
Davidson, D. 1994b. “The social aspect of language.” In *The Philosophy of Michael Dummett*, B. McGuinness and G. Oliveri (eds), 1–16. Dordrecht: Kluwer.
Davidson, D. 1995. “Law and cause.” *Dialectica* 49: 263–279.

Davidson, D. 1997. "Indeterminism and antirealism." In *Realism/Antirealism and Epistemology*, C. B. Kulp (ed.), 109–122. Lanham, MD: Rowman & Littlefield.

Davidson, D. 1998. "Replies." *Crítica* 30: 97–112.

Davidson, D. 2001a. "Appendix to essay 10." In *Inquiries intro Truth and Interpretation. Second Edition*, D. Davidson, 281–285. Oxford: Clarendon Press.

Davidson, D. 2001b. "Comments on Karlovy Vary papers." In *Interpreting Davidson*, P. Kotatko, P. Pagin and G. Segal (eds), 285–307. Stanford: CSLI Publications.

Davidson, D. 2001c. *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press.

Davidson, D. 2001d. "What thought requires." In *The Foundations of Cognitive Science*, J. Branquinho (ed.), 121–132. Oxford: Clarendon Press.

Davies, M. 1995. "Philosophy of mind." In *Philosophy: A Guide through the Subject*, A. C. Grayling (ed.), 250–335. Oxford: Oxford University Press.

Davies, M. and Stone, T. (eds.). 1995. *Folk Psychology. The Theory of Mind Debate.* Oxford: Blackwell.

DeBellis, M. 1995. *Music and Conceptualization*. Cambridge: Cambridge University Press.

Dennett, D. C. 1969. *Content and Consciousness*. London: Routledge and Kegan Paul.

Dennett, D. C. 1971. "Intentional systems." *Journal of Philosophy* 68: 87–106.

Dennett, D. C. 1978. *Brainstorms*. Montgomery, VT: Bradford Books.

Dennett, D. C. 1982. "Comments on Rorty." *Synthese* 53: 349–356.

Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. C. 1988a. "Précis of *The Intentional Stance*". *Behavioral and Brain Sciences* 11: 495–505.

Dennett, D. C. 1988b. "Quining qualia." In *Consciousness in Contemporary Science*, A. J. Marcel and E. Bisiach (eds), 42–77. New York: Oxford University Press.

Dennett, D. C. 1988c. "Science, philosophy, and interpretation." *Behavioral and Brain Sciences* 11: 535–546.

Dennett, D. C. 1990a. "The myth of original intentionality." In *Modelling the Mind*, K. A. Mohyeldin Said, W. H. Newton-Smith, R. Viale and K. V. Wilkes (eds), 43–62. Oxford: Clarendon Press.

Dennett, D. C. 1990b. "The interpretation of texts, people and other artifacts." *Philosophy and Phenomenological Research* 50: 177–194.

Dennett, D. C. 1991a. *Consciousness Explained*. Boston, MA: Little, Brown & Co.

Dennett, D. C. 1991b. "Real patterns." *Journal of Philosophy* 87: 27–51.

Dennett, D. C. 1993a. "Back from the drawing board." In *Dennett and His Critics. Demystifying Mind*, B. Dahlbom (ed.), 203–235. Oxford: Blackwell.

Dennett, D. C. 1993b. "The message is: there is no medium." *Philosophy and Phenomenological Research* 53: 919–931.

Dennett, D. C. 1994. "Get real." *Philosophical Topics* 22: 505–568.

Dennett, D. C. 1995. *Darwin's Dangerous Idea*. New York, NY: Simon & Schuster.

Dennett, D. C. 1996. *Kinds of Minds*. London: Weidenfeld and Nicolson.

Dennett, D. C. 1998. *Brainchildren: Essays on Designing Minds*. Cambridge, MA: MIT Press.

Dennett, D. C. 2000. "With a little help from my friends." In *Dennett's Philosophy: A Comprehensive Assessment*, D. Ross, A. Brook and D. Thompson (eds), 327–388. Cambridge, MA: MIT Press.

Doppelt, G. 1978. "Incorrigibility and the mental." *Australasian Journal of Philosophy* 56: 3–20.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Oxford: Basil Blackwell.

Dretske, F. 1993. "Conscious experience." *Mind* 102: 1–21.
Dretske, F. 1995. "Phenomenal externalism or if meanings ain't in the head, where are qualia?" *Philosophical Issues, Vol. 7: Perception*, 143–158.
Dummett, M. 1959. "Truth." *Proceedings of the Aristotelian Society* 59: 141–162.
Dummett , M. 1978. *Truth and Other Enigmas*. London: Duckworth.
Ehring, D. 1996. "Mental causation, determinables and property instances." *Noûs* 30: 461–480.
Ellis, B. 2001. *Scientific Essentialism*. Cambridge: Cambridge University Press.
Ellis, B. 2002. *The Philosophy of Nature: A Guide to the New Essentialism*. Chesham: Acumen.
Elton, M. 2003. *Daniel Dennett: Reconciling Science and Our Self-Conception*. Cambridge: Polity Press.
Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.
Evnine, S. 1991. *Donald Davidson*. Oxford: Polity Press.
Feigl, H. 1958. "The 'mental' and the 'physical'." *Minnesota Studies in the Philosophy of Science* 2: 370–497.
Field, H. 1980. *Science without Numbers*. Princeton: Princeton University Press.
Fodor, J. A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.
Fodor, J. A. 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press.
Fodor, J. A. 1985. "Fodor's guide to mental representations." *Mind* 94: 76–100.
Fodor, J. A. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.
Fodor, J. A. 1990. "Making mind matter more." In *A Theory of Content and Other Essays*, J. A. Fodor, 137–159. Cambridge, MA: MIT Press.
Fodor, J. A. and Lepore, E. 1992. *Holism: A Shopper's Guide*. Oxford: Blackwell.
Fodor, J. A. and Lepore, E. 1993. "Is intentional ascription intrinsically normative?" In *Dennett and His Critics. Demystifying Mind*, B. Dahlbom (ed.), 70–82. Oxford: Blackwell.
Føllesdal, D. 1982. "The status of rationality assumptions in interpretation and in the explanation of action." *Dialectica* 36: 301–316.
Føllesdal, D. 1990. "Indeterminacy and mental states." In *Perspectives on Quine*, R. Barrett and R. Gibson (eds), 98–109. Oxford: Oxford University Press.
Frankish, K. 2004. *Mind and Supermind*. Cambridge: Cambridge University Press.
Gauker, C. 1986. "The principle of charity." *Synthese* 69: 1–25.
Gauker, C. 1988. "Objective interpretationism." *Pacific Philosophical Quarterly* 69: 136–151.
Gauker, C. 1994. *Thinking Out Loud: An Essay on the Relation between Thought and Language*. Princeton, NJ: Princeton University Press.
Gauker, C. 2003a. "Attitudes without psychology." *Facta Philosophica* 5: 239–256.
Gauker, C. 2003b. *Words without Meaning*. Cambridge, MA: MIT Press.
Goldman, A. 1989. "Interpretation psychologized." *Mind and Language* 4: 161–185.
Goldman, A. 1993. "The psychology of folk psychology." *Behavioral and Brain Sciences* 16: 15–28.
Goldman, A. 2002. "Simulation theory and mental concepts." In *Simulation and Knowledge of Action*, J. Dokic and J. Proust (eds), 1–19. Amsterdam/Philadelphia: John Benjamins.
Goldberg, S. and Pessin, A. 1997. *Gray Matters. An Introduction to the Philosophy of Mind*. New York, NY: M.E. Sharpe.
Grandy, R. 1973. "Reference, meaning and belief." *Journal of Philosophy* 70: 439–452
Grice, H. P. 1957. "Meaning." *The Philosophical Review* 66: 377–388.
Hacking, I. 1986. "The parody of conversation." In *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, E. Lepore (ed.), 447–458. Oxford: Basil Blackwell.
Haugeland, J. 1982. "The mother of intention." *Nous* 16: 613–619.

Haugeland, J. 1993. "Pattern and being." In *Dennett and His Critics. Demystifying Mind*, B. Dahlbom (ed.), 53–69. Oxford: Blackwell.
Heal, J. 1997. "Radical interpretation." In *A Companion to the Philosophy of Language*, B. Hale and C. Wright (eds), 175–196. Oxford: Blackwell.
Heal, J. 1998. "Co-cognition and off-line simulation: two ways of understanding the simulation approach." *Mind and Language* 13: 477–498.
Heal, J. 2003. "What are psychological concepts for?" In *Mind, Reason and Imagination*, J. Heal, 225–249. Cambridge: Cambridge University Press.
Heck, R. 2000. "Nonconceptual content and the 'space of reasons'." *Philosophical Review* 109: 483–523.
Heil, J. 1988. "Privileged access." *Mind* 97: 238–251.
Heil, J. 1992. *The Nature of True Minds*. Cambridge: Cambridge University Press.
Heil, J. 1998a. *Philosophy of Mind: A Contemporary Introduction*. London: Routledge.
Heil, J. 1998b. "Supervenience deconstructed." *European Journal of Philosophy* 6: 146–155.
Heil, J. 1999. "Multiple realizability." *American Philosophical Quarterly* 36: 189–208.
Heil, J. 2002. "Functionalism, realism and levels of being." In *Hilary Putnam: Pragmatism and Realism*, J. Conant and U. M. Żegleń (eds), 128–142. London: Routledge.
Heil, J. 2003a. *From an Ontological Point of View*. Oxford: Clarendon Press.
Heil, J. 2003b. "Levels of reality." *Ratio* 16: 205–221.
Heil, J. 2003c. "Multiply realized properties." In *Physicalism and Mental Causation*, S. Walter and H.-D. Heckmann (eds), 11–30. Shöningh: Imprint Academic.
Heil, J. and Robb, D. 2003. "Mental properties." *American Philosophical Quarterly* 40: 175–191.
Heitner, R. M. 2000. "Is design relative of real? Dennett on intentional relativism and physical realism." *Minds and Machines* 10: 267–275.
Hempel, C. G. 1980. "Comments on Goodman's *Ways of Worldmaking*." *Synthese* 45: 193–199.
Henderson, D. K. 1987. "The principle of charity and the problem of irrationality (translation and the problem of irrationality)." *Synthese* 73: 225–252.
Holton, R. 1993. "Intention detecting." *The Philosophical Quarterly* 43: 298–318.
Honderich, T. 1982. "The argument for anomalous monism." *Analysis* 42: 59–64.
Hooker, C. A. 1981. "Towards a general theory of reduction." *Dialogue* 20: 38–59; 201–236; 496–529.
Horgan, T. 1993. "The austere ideology of folk psychology." *Mind and Language* 8: 282–297.
Hornsby, J. 1997. *Simple Mindedeness: In Defence of Naïve Naturalism in the Philosophy of Mind*. Cambridge, MA: Harvard University Press.
Horwich, P. 1998a. *Meaning*. Oxford: Clarendon Press.
Horwich, P. 1998b. *Truth. 2nd edition*. Oxford: Clarendon Press.
Horwich, P. 2005. *Reflections on Meaning*. Oxford: Clarendon Press.
Hurley, S. 2001. "Overintellectualizing the mind." *Philosophy and Phenomenological Research* 63: 423–431.
Hutto, D. 1999. *The Presence of Mind*. Amsterdam/Philadelphia: John Benjamins.
Hüttemann, A. and Papineau, D. 2005. "Physicalism decomposed." *Analysis* 65: 33–39.
Jacob, P. 2002. "Some problems for reductive physicalism." *Philosophy and Phenomenological Research* 65: 648–654.
Jackson, F. 1981. "On property identity." *Philosophia* 11: 289–305.
Jackson, F. 1993. "Appendix A (for philosophers)." *Philosophy and Phenomenological Research* 53: 899–903.

Jackson, F. 1995. "Essentialism, mental properties and causation." *Proceedings of the Aristotelian Society* 95: 253–268.
Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon Press.
Jackson, F. 2002. "From reduction to type-type identity." *Philosophy and Phenomenological Research* 65: 644–647.
Jackson, F. 2003. "From H_2O to water: the relevance to a priori passage." In *Real Metaphysics. Essays in Honour of D. H. Mellor*, H. Lillehammer and G. Rodriguez-Pereyra (eds), 84–97. London, New York: Routledge.
Jackson, F. and Pettit, P. 1988. "Functionalism and broad content." *Mind* 97: 381–400.
Jackson, F. and Pettit, P. 1990a. "Causation in the philosophy of mind." *Philosophy and Phenomenological Research* 50: 195–214.
Jackson, F. and Pettit, P. 1990b. "Program explanation: a general perspective." *Analysis* 50: 107–117.
Jackson, F. and Pettit, P. 1993. "Folk belief and commonplace belief." *Mind and Language* 8: 298–305.
Jakab, Z. 2000. "Ineffability of qualia: a straightforward naturalistic explanation." *Consciousness and Cognition* 9: 329–351.
Johnston, M. 1992. "How to speak of the colors." *Philosophical Studies* 68: 221–264.
Jeffrey, R. 1983. *The Logic of Decision. 2nd ed.* Chicago, IL: University of Chicago Press.
Kemmerling, A. 1993. "The philosophical significance of a shared language." In *Reflecting Davidson*, R. Stoecker (ed.), 85–116. Berlin: Walter de Gruyter.
Kim, J. 1976. "Events as property exemplifications." In *Action Theory*, M. Brand, D. Walton (eds), 159–177. Dordrecht: Reidel.
Kim, J. 1984a. "Concepts of supervenience." *Philosophy and Phenomenological Research* 45: 153–176.
Kim, J. 1984b. "Epiphenomenal and supervenient causation." *Midwest Studies in Philosophy* 9: 257–270.
Kim, J. 1992. "Multiple realization and the metaphysics of reduction." *Philosophy and Phenomenological Research* 52: 1–26.
Kim, J. 1993a. "Can supervenience and 'non-strict laws' save anomalous monism?" In *Mental causation*, John Heil and Alfred Mele (eds), 19–26. Oxford: Clarendon Press.
Kim, J. 1993b. "Postscripts on mental causation." In *Supervenience and Mind*, J. Kim, 358–367. Cambridge: Cambridge University Press.
Kim, J. 1993c. "Postscripts on supervenience." In *Supervenience and Mind*, J. Kim, 161–171. Cambridge: Cambridge University Press.
Kim, J. 1993d. "The non-reductivist's troubles with mental causation." In *Mental Causation*, J. Heil and A. Mele (eds), 189–210. Oxford: Clarendon Press.
Kim, J. 1994. "The myth of nonreductive materialism." In *The Mind-Body Problem*, R. Warner and T. Szubka (eds), 242–260. Oxford: Blackwell.
Kim, J. 1995. "Mental causation: What? Me worry?" *Philosophical Issues* 6: 123–151.
Kim, J. 1996. *Philosophy of Mind*. Boulder, CO: Westview Press.
Kim, J. 1998. *Mind in a Physical World. An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.
Kim, J. 2002. "Replies." *Philosophy and Phenomenological Research* 65: 671–680.

Kim, J. "2003a." Blocking causal drainage and other maintenance chores with mental causation." *Philosophy and Phenomenological Research* 67: 151–176.
Kim, J. 2003b. "Supevenience, emergence, realization, reduction." In *The Oxford Handbook of Metaphysics*, Michael J. Loux and Dean W. Zimmerman (eds), 556–584. Oxford: Oxford University Press.
Kim, J. 2005. *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.
Kirk, R. 1999. *Relativism and Reality: A Contemporary Introduction*. London & New York: Routledge.
Klagge, J. 1988. "Supervenience: ontological and ascriptive." *Australasian Journal of Philosophy* 66: 461–70.
Kripke, S. 1979. "A puzzle about belief." In *Meaning and Use*, A. Margalit (ed.), 239–283. Dordrecht: D. Reidel.
Kripke, S. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Lance, M. and O'Leary-Hawthorne, J. 1997. *The Grammar of Meaning: Normativity and Semantic Discourse*. Cambridge: Cambridge University Press.
Laurier, D. 1999. "On the principle of charity and the sources of indeterminacy." In *Consciousness and Intentionality: Models and Modalities of Attribution*, D. Fisette (ed.), 229–248. Dordrecht: Kluwer.
Lepore, E. and Ludwig, K. 2005. *Donald Davidson: Meaning, Truth, Language, and Reality*. Oxford: Clarendon Press.
Levin, J. 1988. "Must reasons be rational?" *Philosophy of Science* 55: 199–217.
Lewis, D. K. 1966. "An argument for the identity theory." *Journal of Philosophy* 63: 17–25.
Lewis, D. K. 1970. "How to define theoretical terms." *Journal of Philosophy* 67: 427–446.
Lewis, D. K. 1972. "Psychophysical and theoretical identifications". *Australasian Journal of Philosophy* 50: 249–258.
Lewis, D. K. 1974. "Radical interpretation." *Synthese* 23: 331–344.
Lewis, D. K. 1980. "Mad pain and Martian pain." In *Readings in the Philosophy of Psychology*, vol. 1, Ned Block (ed.), 216–222. Cambridge, MA: Harvard University Press.
Lewis, D. K. 1983a. "New work for a theory of universals." *Australasian Journal of Philosophy* 61: 343–377.
Lewis, D. K. 1983b. *Philosophical Papers, vol. 1*. Oxford: Oxford University Press.
Lewis, D. K. 1994. "Reduction of mind." In *A Companion to the Philosophy of Mind*, S. Guttenplan (ed.), 412–431. Oxford: Blackwell.
Lloyd, D. 2000. "Popping the thought balloon." In *Dennett's Philosophy: A Comprehensive Assessment*, D. Ross, A. Brook and D. Thompson (eds), 169–199. Cambridge, MA: MIT Press.
Ludwig, K. 2003. "The mind-body problem: an overview." In *The Blackwell Guide to Philosophy of Mind*, S. Stich and T. Warfield (eds), 1–46. Oxford: Blackwell.
Lycan, W. 1988. *Judgement and Justification*. Cambridge: Cambridge University Press.
Lyons, W. 1995. *Approaches to Intentionality*. Oxford: Clarendon Press.
Macdonald, C. 2005. *Varieties of Things: Foundations of Contemporary Metaphysics*. Oxford: Blackwell.
Manning, R. N. 2003. "Intepretation, reasons, and facts." *Inquiry* 46: 346–376.
Maund, B. 2003. *Perception*. Chesham: Acumen Press.
McCulloch, G. 1990. "Dennett's little grains of salt." *Philosophical Quarterly* 40: 1–12.
McDowell, J. 1994. *Mind and World*. Cambridge, MA: Harvard University Press.

McGinn, C. 1978. "Mental states, natural kinds and psychophysical laws." *Proceedings of the Aristotelian Society, Suppl. Vol.* 2: 195–220.

McLaughlin, B. 1993. "On Davidson's response to the charge of epiphenomenalism." In *Mental Causation*, J. Heil and A. Mele (eds), 27–40. Oxford: Clarendon Press.

McLaughlin, B. and O'Leary-Hawthorne, J. 1994. "Dennett's logical behaviorism." *Philosophical Topics* 22: 189–258.

Mellor, D. H. 1993. "Properties and predicates." In *Ontology, Causality and Mind: Essays in Honour of D. M. Armstrong*, J. Bacon, K. Campbell, and L. Reinhardt (eds), 101–113. Cambridge: Cambridge University Press.

Mellor, D. H. 1995. *The Facts of Causation*. London: Routledge.

Metzinger, T. and Walde, B. 2000. "Commentary on Jakab's 'Ineffability of qualia'." *Consciousness and Cognition* 9: 352–362.

Millar, A. 2004. *Understanding People: Normativity and Rationalizing Explanation*. Oxford: Clarendon Press.

Miller, A. 1998. *Philosophy of Language*. London: UCL Press.

Moran, R. 1994. "Interpretation theory and the first person." *Philosophical Quarterly* 44: 154–173.

Moran, R. 2001. *Authority and Estrangement. An Essay on Self-Knowledge*. Princeton, NJ: Princeton University Press.

Morton, A. 2003a. "Saving belief from (internalist) epistemology." *Facta Philosophica* 5: 277–295.

Morton, A. 2003b. *The Importance of Being Understood: Folk Psychology as Ethics*. London: Routledge.

Mölder, B. 1997. "First-person authority about one's own mind." *Trames* 1: 50–64.

Mölder, B. 1998. "Vaimu tõeväärtustamine." In *Arutlusi teadusest ja tõest*, J. Kangilaski and E. Kasak (eds), 90–95. (English abstract: "Attribution of truth-values to mind", 165–166.) Võru: Võru Instituut.

Nagel, E. 1961. *The Structure of Science*. New York, NY: Harcourt, Brace & World.

Nichols, S. and Stich, S. 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Clarendon Press.

Noordhof, P. 1998. "Do tropes resolve the problem of mental causation?" *Philosophical Quarterly* 48: 221–226.

Norton, J. D. 2003. "Causation as folk science." *Philosophers' Imprint* 3 (4), 1–22.

Olson, D. R. and Kamawar, D. 1999. "The theory of ascriptions." In *Developing Theories of Intention: Social Understanding and Self-Control*, P. D. Zelazo, J. W. Astington and D. R. Olson (eds), 153–166. Mahwah, NJ: Lawrence Erlbaum.

Peacocke, C. 1983. *Sense and Content: Experience, Thought, and their Relations*. Oxford: Clarendon Press.

Peacocke, C. 1986. *Thoughts: An Essay on Content*. Oxford: Basil Blackwell.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Peacocke, C. 1994. "Nonconceptual content: kinds, rationales, and relations." *Mind and Language* 9: 419–429.

Peacocke, C. 2001. "Does perception have a nonconceptual content." *Journal of Philosophy* 98: 239–264.

Pereboom, D. and Kornblith, H. 1991. "The metaphysics of irreducibility." *Philosophical Studies* 63: 125–145.

Place, U. 1956. "Is consciousness a brain process." *British Journal of Psychology* 47: 44–50.

Putnam, H. 1967. “Psychological Predicates.” In *Art, Mind and Religion*, W. H. Capitan and D. D. Merrill (eds), 37–48. Pittsburgh, PA: University of Pittsburgh Press.

Putnam, H. 1969. “On properties.” In *Essays in Honor of Carl G. Hempel*, N. Rescher (ed.), 235–254. Dordrecht: D. Reidel.

Putnam, H. 1975a. *Mind, Language, and Reality. Philosophical Papers, vol. 2*. Cambridge: Cambridge University Press.

Putnam, H. 1975b. “Philosophy and our mental life.” In *Mind, Language, and Reality. Philosophical Papers, vol. 2*, H. Putnam, 291–303. Cambridge: Cambridge University Press.

Putnam, H. 1975c. “The meaning of ‘meaning’.” *Minnesota Studies in the Philosophy of Science* 7: 131–193.

Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Quine, W. V. O. 1981. *Theories and Things*. Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1987. “Indeterminacy of translation again.” *Journal of Philosophy* 84: 5–10.

Rawling, P. 2003. “Radical interpretation.” In *Donald Davidson*, K. Ludwig (ed.), 85–112. Cambridge: Cambridge University Press.

Rey, G. 1994. “Dennett’s unrealistic psychology.” *Philosophical Topics* 22: 259–289.

Rey, G. 1997. *Contemporary Philosophy of Mind*. Oxford: Blackwell.

Robb, D. 1997. “The properties of mental causation.” *Philosophical Quarterly* 47: 178–194.

Rosenthal, D. 1994. “First-person operationalism and mental taxonomy.” *Philosophical Topics* 22: 319–349.

Russell, B. 1913. “On the notion of cause.” *Proceedings of the Aristotelian Society* 13: 1–26.

Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson.

Schaffner, K. 1967. “Approaches to reduction.” *Philosophy of Science* 34: 137–147.

Schiffer, S. 1987. *Remnants of Meaning*. Cambridge, MA: MIT Press.

Schiffer, S. 2003. *The Things We Mean*. Oxford: Clarendon Press.

Seager, W. 2000. “Real patterns and surface metaphysics.” In *Dennett’s Philosophy: A Comprehensive Assessment*, D. Ross, A. Brook and D. Thompson (eds), 95–129. Cambridge, MA: MIT Press.

Searle, J. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Sedivy, S. 2004. “Minds: contents without vehicles.” *Philosophical Psychology* 17: 149–179.

Segal, G. 2000. *A Slim Book about Narrow Content*. Cambridge, MA: MIT Press.

Sellars, W. 1956. “Empiricism and the philosophy of mind.” *Minnesota Studies in the Philosophy of Science* 1: 253–329.

Shoemaker, S. 1980. “Causality and properties.” In *Time and Cause*, P. Van Inwagen (ed.), 109–135. Dordrecht: Reidel.

Shoemaker, S. 1981. “Some varieties of functionalism.” *Philosophical Topics* 12: 83–118.

Shoemaker, S. 1994. “Self-knowledge and ‘inner sense’.” *Philosophy and Phenomenological Research* 54: 249–314.

Shoemaker, S. 2001. “Realization and mental causation.” In *Physicalism and Its Discontents*, C. Gillett and B. Loewer (eds), 74–98. Cambridge: Cambridge University Press.

Schulz, S. 2002. *Alien Minds: Investigating Eliminative Materialism*. Paderborn: Mentis.

Sklar, L. 1967. “Types of inter-theoretic reduction.” *British Journal for the Philosophy of Science* 18: 109–124.

Smart, J. 1959. “Sensations and brain processes.” *Philosophical Review* 68: 141–156.

Smedslund, J. 1993. “How shall the concept of anger be defined?” *Theory & Psychology* 3: 5–33.

Smedslund, J. 1997. *The Structure of Psychological Common Sense*. Mahwah, NJ: Lawrence Erlbaum.
Sosa, E. 1984. "Mind-body interaction and supervenient causation." *Midwest studies in philosophy* 9: 271–282.
Spohn, W. 1998. "How to understand the foundations of empirical belief in a coherentist way." *Proceedings of the Aristotelian Society* 98: 23–40.
Spohn, W. 1999. "Two coherence principles." *Erkenntnis* 50: 155–175.
Spohn, W. 2006. "Causation: an alternative." *British Journal for the Philosophy of Science* 57: 93–119.
Steward, H. 1997. *The Ontology of Mind: Events, Processes and States*. Oxford: Clarendon Press.
Stich, S. 1981. "Dennett on intentional systems." *Philosophical Topics* 12: 39–61.
Stich, S. 1983. *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.
Stich, S. 1990. *The Fragmentation of Reason*. Cambridge, MA: MIT Press.
Stich, S. 1996. *Deconstructing the Mind*. Cambridge, MA: MIT Press.
Stoljar, D. 2001. "Physicalism." *The Stanford Encyclopedia of Philosophy* (Spring 2001 Edition), E. N. Zalta (ed.).
Strawson, P. 1985. "Causation and explanation." In *Essays on Davidson: Actions and Events*, B. Vermazen and M. Hintikka (eds), 115–135. Oxford: Oxford University Press.
Szabo Gendler, T. and Hawthorne, J. (eds.). 2006. *Perceptual Experience*. Oxford: Oxford University Press.
Tanney, J. 2002. "Self-knowledge, normativity, and construction." In *Logic, Thought and Language*, A. O'Hear (ed.), 37–55. Cambridge: Cambridge University Press.
Thagard, P. and Kunda, Z. 1998. "Making sense of people: coherence mechanisms." In *Connectionist Models of Social Reasoning and Social Behavior*, S. J. Read and L. C. Miller (eds), 3–26. Hillsdale, NJ: Erlbaum.
Thagard, P. and Nisbett, R. E. 1983. "Rationality and charity." *Philosophy of Science* 50: 250–267.
Thagard, P. 1992. *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.
Thagard, P. 2000. *Coherence in Thought and Action*. Cambridge, MA: MIT Press.
Van Gelder, T. 1998. "Monism, dualism, pluralism." *Mind and Language* 13: 76–97.
Van Gulick, R. 1992. "Three bad arguments for intentional property epiphenomenalism." *Erkenntnis* 36: 311–332.
Viger, C. 2000. "Where do Dennett's stances stand? Explaining our kind of mind." In *Dennett's Philosophy: A Comprehensive Assessment*, D. Ross, A. Brook and D. Thompson (eds), 131–145. Cambridge, MA: MIT Press.
Wilson, N. 1959. "Substances without substrata." *Review of Metaphysics* 12: 529–539.
Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.
Wright, C. 1987. "On making up one's mind: Wittgenstein on intention." In *Logic, Philosophy of Science and Epistemology*, P. Weingartner and G. Schurz (eds), 391–404. Vienna: Holder-Pickler-Tempsky.
Wright, C. 1989a. "Wittgenstein's later philosophy of mind: sensation, privacy and intention." *Journal of Philosophy* 76: 622–634.
Wright, C. 1989b. "Wittgenstein's rule-following considerations and the central project of theoretical linguistics." In *Reflections on Chomsky*, A. George (ed.), 233–264. Oxford: Blackwell.
Wright, C. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Wright, C. 1993. "Realism: The contemporary debate – whither now?" In *Reality: Representation and Projection*, J. Haldane and C. Wright (eds), 63–84. Oxford: Oxford University Press.
Wright, C. 1998. "Self-knowledge: the Wittgensteinian legacy." In *Current Issues in Philosophy of Mind*, A. O'Hear (ed.), 101–122. Cambridge: Cambridge University Press.
Wright, C. 2002. "What could antirealism about ordinary psychology possibly be?" *The Philosophical Review* 111: 205–233.
Yablo, S. 1992. "Mental causation." *The Philosophical Review* 101: 245–280.

Index

Advances in Consciousness Research

A complete list of titles in this series can be found on the publishers' website, *www.benjamins.com*

80 **MÖLDER, Bruno:** Mind Ascribed. An elaboration and defence of interpretivism. 2010. xii, 293 pp.

79 **PERRY, Elaine, Daniel COLLERTON, Fiona E.N. LEBEAU and Heather ASHTON (eds.):** New Horizons in the Neuroscience of Consciousness. xxv, 323 pp. + index. *Expected September 2010*

78 **CZIGLER, István and István WINKLER (eds.):** Unconscious Memory Representations in Perception. Processes and mechanisms in the brain. 2010. x, 274 pp.

77 **GLOBUS, Gordon G.:** The Transparent Becoming of World. A crossing between process philosophy and quantum neurophilosophy. 2009. xiii, 169 pp.

76 **BRÅTEN, Stein:** The Intersubjective Mirror in Infant Learning and Evolution of Speech. 2009. xxii, 351 pp.

75 **SKRBINA, David (ed.):** Mind that Abides. Panpsychism in the new millennium. 2009. xiv, 401 pp.

74 **CAÑAMERO, Lola and Ruth AYLETT (eds.):** Animating Expressive Characters for Social Interaction. 2008. xxiii, 296 pp.

73 **HARDCASTLE, Valerie Gray:** Constructing the Self. 2008. xi, 186 pp.

72 **JANZEN, Greg:** The Reflexive Nature of Consciousness. 2008. vii, 186 pp.

71 **KROIS, John Michael, Mats ROSENGREN, Angela STEIDELE and Dirk WESTERKAMP (eds.):** Embodiment in Cognition and Culture. 2007. xxii, 304 pp.

70 **RAKOVER, Sam S.:** To Understand a Cat. Methodology and philosophy. 2007. xviii, 253 pp.

69 **KUCZYNSKI, John-Michael:** Conceptual Atomism and the Computational Theory of Mind. A defense of content-internalism and semantic externalism. 2007. x, 524 pp.

68 **BRÅTEN, Stein (ed.):** On Being Moved. From mirror neurons to empathy. 2007. x, 333 pp.

67 **ALBERTAZZI, Liliana (ed.):** Visual Thought. The depictive space of perception. 2006. xii, 380 pp.

66 **VECCHI, Tomaso and Gabriella BOTTINI (eds.):** Imagery and Spatial Cognition. Methods, models and cognitive assessment. 2006. xiv, 436 pp.

65 **SHAUMYAN, Sebastian:** Signs, Mind, and Reality. A theory of language as the folk model of the world. 2006. xxvii, 315 pp.

64 **HURLBURT, Russell T. and Christopher L. HEAVEY:** Exploring Inner Experience. The descriptive experience sampling method. 2006. xii, 276 pp.

63 **BARTSCH, Renate:** Memory and Understanding. Concept formation in Proust's *A la recherche du temps perdu*. 2005. x, 160 pp.

62 **DE PREESTER, Helena and Veroniek KNOCKAERT (eds.):** Body Image and Body Schema. Interdisciplinary perspectives on the body. 2005. x, 346 pp.

61 **ELLIS, Ralph D.:** Curious Emotions. Roots of consciousness and personality in motivated action. 2005. viii, 240 pp.

60 **DIETRICH, Eric and Valerie Gray HARDCASTLE:** Sisyphus's Boulder. Consciousness and the limits of the knowable. 2005. xii, 136 pp.

59 **ZAHAVI, Dan, Thor GRÜNBAUM and Josef PARNAS (eds.):** The Structure and Development of Self-Consciousness. Interdisciplinary perspectives. 2004. xiv, 162 pp.

58 **GLOBUS, Gordon G., Karl H. PRIBRAM and Giuseppe VITIELLO (eds.):** Brain and Being. At the boundary between science, philosophy, language and arts. 2004. xii, 350 pp.

57 **WILDGEN, Wolfgang:** The Evolution of Human Language. Scenarios, principles, and cultural dynamics. 2004. xii, 240 pp.

56 **GENNARO, Rocco J. (ed.):** Higher-Order Theories of Consciousness. An Anthology. 2004. xii, 371 pp.

55 **PERUZZI, Alberto (ed.):** Mind and Causality. 2004. xiv, 235 pp.

54 **BEAUREGARD, Mario (ed.):** Consciousness, Emotional Self-Regulation and the Brain. 2004. xii, 294 pp.

53 **HATWELL, Yvette, Arlette STRERI and Edouard GENTAZ (eds.):** Touching for Knowing. Cognitive psychology of haptic manual perception. 2003. x, 322 pp.

52 **NORTHOFF, Georg:** Philosophy of the Brain. The brain problem. 2004. x, 433 pp.

51 **DROEGE, Paula:** Caging the Beast. A theory of sensory consciousness. 2003. x, 183 pp.

50 **GLOBUS, Gordon G.:** Quantum Closures and Disclosures. Thinking-together postphenomenology and quantum brain dynamics. 2003. xxii, 200 pp.

49 **OSAKA, Naoyuki (ed.):** Neural Basis of Consciousness. 2003. viii, 227 pp.

48 **JIMÉNEZ, Luis (ed.):** Attention and Implicit Learning. 2003. x, 385 pp.

47 **COOK, Norman D.:** Tone of Voice and Mind. The connections between intonation, emotion, cognition and consciousness. 2002. x, 293 pp.

46 **MATEAS, Michael and Phoebe SENGERS (eds.):** Narrative Intelligence. 2003. viii, 342 pp.

45 **DOKIC, Jérôme and Joëlle PROUST (eds.):** Simulation and Knowledge of Action. 2002. xxii, 271 pp.

44 **MOORE, Simon C. and Mike OAKSFORD (eds.):** Emotional Cognition. From brain to behaviour. 2002. vi, 350 pp.

43 **DEPRAZ, Nathalie, Francisco J. VARELA and Pierre VERMERSCH:** On Becoming Aware. A pragmatics of experiencing. 2003. viii, 283 pp.

42 **STAMENOV, Maxim I. and Vittorio GALLESE (eds.):** Mirror Neurons and the Evolution of Brain and Language. 2002. viii, 392 pp.

41 **ALBERTAZZI, Liliana (ed.):** Unfolding Perceptual Continua. 2002. vi, 296 pp.

40 **MANDLER, George:** Consciousness Recovered. Psychological functions and origins of conscious thought. 2002. xii, 142 pp.

39 **BARTSCH, Renate:** Consciousness Emerging. The dynamics of perception, imagination, action, memory, thought, and language. 2002. x, 258 pp.

38 **SALZARULO, Piero and Gianluca FICCA (eds.):** Awakening and Sleep–Wake Cycle Across Development. 2002. vi, 283 pp.

37 **PYLKKÄNEN, Paavo and Tere VADÉN (eds.):** Dimensions of Conscious Experience. 2001. xiv, 209 pp.

36 **PERRY, Elaine, Heather ASHTON and Allan H. YOUNG (eds.):** Neurochemistry of Consciousness. Neurotransmitters in mind. With a foreword by Susan Greenfield. 2002. xii, 344 pp.

35 **Mc KEVITT, Paul, Seán Ó NUALLÁIN and Conn MULVIHILL (eds.):** Language, Vision and Music. Selected papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing, Galway, 1999. 2002. xii, 433 pp.

34 **FETZER, James H. (ed.):** Consciousness Evolving. 2002. xx, 253 pp.

33 **YASUE, Kunio, Mari JIBU and Tarcisio DELLA SENTA (eds.):** No Matter, Never Mind. Proceedings of Toward a Science of Consciousness: Fundamental approaches, Tokyo 1999. 2002. xvi, 391 pp.

32 **VITIELLO, Giuseppe:** My Double Unveiled. The dissipative quantum model of brain. 2001. xvi, 163 pp.

31 **RAKOVER, Sam S. and Baruch CAHLON:** Face Recognition. Cognitive and computational processes. 2001. x, 306 pp.

30 **BROOK, Andrew and Richard C. DEVIDI (eds.):** Self-Reference and Self-Awareness. 2001. viii, 277 pp.

29 **VAN LOOCKE, Philip (ed.):** The Physical Nature of Consciousness. 2001. viii, 321 pp.

28 **ZACHAR, Peter:** Psychological Concepts and Biological Psychiatry. A philosophical analysis. 2000. xx, 342 pp.

27 **GILLETT, Grant R. and John McMILLAN:** Consciousness and Intentionality. 2001. x, 265 pp.

26 **Ó NUALLÁIN, Seán (ed.):** Spatial Cognition. Foundations and applications. 2000. xvi, 366 pp.

25 **BACHMANN, Talis:** Microgenetic Approach to the Conscious Mind. 2000. xiv, 300 pp.

24 **ROVEE-COLLIER, Carolyn, Harlene HAYNE and Michael COLOMBO:** The Development of Implicit and Explicit Memory. 2000. x, 324 pp.

23 **ZAHAVI, Dan (ed.):** Exploring the Self. Philosophical and psychopathological perspectives on self-experience. 2000. viii, 301 pp.

22 **ROSSETTI, Yves and Antti REVONSUO (eds.):** Beyond Dissociation. Interaction between dissociated implicit and explicit processing. 2000. x, 372 pp.

21 **HUTTO, Daniel D.:** Beyond Physicalism. 2000. xvi, 306 pp.

20 **KUNZENDORF, Robert G. and Benjamin WALLACE (eds.):** Individual Differences in Conscious Experience. 2000. xii, 412 pp.

19 **DAUTENHAHN, Kerstin (ed.):** Human Cognition and Social Agent Technology. 2000. xxiv, 448 pp.

18 **PALMER, Gary B. and Debra J. OCCHI (eds.):** Languages of Sentiment. Cultural constructions of emotional substrates. 1999. vi, 272 pp.

17 **HUTTO, Daniel D.:** The Presence of Mind. 1999. xiv, 252 pp.

16 **ELLIS, Ralph D. and Natika NEWTON (eds.):** The Caldron of Consciousness. Motivation, affect and self-organization — An anthology. 2000. xxii, 276 pp.

15 **CHALLIS, Bradford H. and Boris M. VELICHKOVSKY (eds.):** Stratification in Cognition and Consciousness. 1999. viii, 293 pp.

14 **SHEETS-JOHNSTONE, Maxine:** The Primacy of Movement. 1999. xxxiv, 583 pp.

13 **VELMANS, Max (ed.):** Investigating Phenomenal Consciousness. New methodologies and maps. 2000. xii, 381 pp.

12 **STAMENOV, Maxim I. (ed.):** Language Structure, Discourse and the Access to Consciousness. 1997. xii, 364 pp.

11 **PYLKKÖ, Pauli:** The Aconceptual Mind. Heideggerian themes in holistic naturalism. 1998. xxvi, 297 pp.

10 **NEWTON, Natika:** Foundations of Understanding. 1996. x, 211 pp.

9 **Ó NUALLÁIN, Seán, Paul Mc KEVITT and Eoghan Mac AOGÁIN (eds.):** Two Sciences of Mind. Readings in cognitive science and consciousness. 1997. xii, 490 pp.

8 **GROSSENBACHER, Peter G. (ed.):** Finding Consciousness in the Brain. A neurocognitive approach. 2001. xvi, 326 pp.

7 **MAC CORMAC, Earl and Maxim I. STAMENOV (eds.):** Fractals of Brain, Fractals of Mind. In search of a symmetry bond. 1996. x, 359 pp.

6 **GENNARO, Rocco J.:** Consciousness and Self-Consciousness. A defense of the higher-order thought theory of consciousness. 1996. x, 220 pp.

5 **STUBENBERG, Leopold:** Consciousness and Qualia. 1998. x, 368 pp.

4 **HARDCASTLE, Valerie Gray:** Locating Consciousness. 1995. xviii, 266 pp.

3 **JIBU, Mari and Kunio YASUE:** Quantum Brain Dynamics and Consciousness. An introduction. 1995. xvi, 244 pp.

2 **ELLIS, Ralph D.:** Questioning Consciousness. The interplay of imagery, cognition, and emotion in the human brain. 1995. viii, 262 pp.

1 **GLOBUS, Gordon G.:** The Postmodern Brain. 1995. xii, 188 pp.